Searching dbSNP for Variations and Variation Data

Created: July 07, 2005
Updated: January 22, 2007

This section contains FAQs that provide information about general dbSNP search options as well as FAQs that cover specific SNP searches — everything from searching for SNPs in populations to searching for SNPs in genes and chromosomes. There are FAQs in this section about searching for SNP primers, and FAQs about searching for a SNP with a specific ID. There are FAQs that show you how to search for specific SNP data, how to conduct Batch queries under various circumstances, and FAQs that show you how to locate SNPs using sequence homology (BLAST). There are even FAQs that show you how to find dbSNP statistical data and data histories or historic data.
To begin searching this section of the dbSNP FAQ Archive, you can either:
Enter your **search word(s)** text in the text box at the top of the left side of the page and click on the “Search” button,
OR
Click on any of the “Searching dbSNP” sub-categories listed in the **navigation box** on the left side of the page to navigate to the sub-category of your choice.

General dbSNP Search Options

Introduction to Searching dbSNP

How do I search dbSNP?

As of April 2006, there are 11 ways to directly search dbSNP:

1. Search by IDConduct your search using the submitted SNP (ss) or refSNP (rs) Id numbers as your search term.
2. Search by submitterConduct your search using the submitter handle, the submitter name or the name of the submitting institution as your search term.
3. Batch QueriesTo retrieve a large number of SNPs in a single batch, you conduct your search using submitted SNP ID numbers (ss#), submitter IDs (handle|local_snp_id), RefSNP cluster ID numbers (rs#), or Entrez SNP results as your search terms.
4. Search by batch (New Batches)Conduct your search by selecting an organism, and batch type. Then supply the submission date, submitter handle and submitter batch ID if available.
5. Search by methodConduct your search using the method type used to identify the SNP as your search term.
6. Search by population classConduct your search using population category information as your search term.
7. Search by population detailConduct your search using the population handle, the submitter population ID, or the population description as your search term.
8. Search by publicationConduct your search using the title of the publication, the author names, or the submitter handle as your search term.
9. Search by locus (gene) conduct your search on Entrez Gene using the gene name or descriptor and the SNP filter: “gene snp”[filter] as your search terms.
10. Search between MarkersConduct your search using STS markers that map to the same chromosome as your search terms.

11. Search mouse SNPs between strainsConduct your search by selecting two mouse strains of interest and limiting the search by functional class, by chromosome, or by Entrez SNP search fields.(**04/05/06**)

Using the Search Filter Options

Do the "Chromosome Position" record field and the "[CHRPOS]" search filter option start at '0' or '1' as the pTel?

The minimum start position is 1. The "[CHRPOS]" search filter option will allow you to get all SNPs having position "0" that don't map to a contig or chromosome. The synonym for "Chromosome Position" used in the Preview/ Index is "Base Position". **(11/30/05)**

What does the [SB] search limit field do?

The [SB] field limits your search to a subset of records containing links (i.e., Snp_protein, snp_nucleotide, etc.).
Aliases=FILT, FLTR, SUBSET, SB, FIL
FullName=Filter
Description=Limits the records

Could I safely use loc_cnt=1 to get uniquely mapped SNPs on the "reference" genome?

No. Use "SNPMapInfo.weight =1 AND SNPMapInfo.ref_cnt > 0" to count SNPs that align at one physical position and also align on the reference genome.

Direct Database Queries

Can dbSNP be queried over the Internet using SQL? I'd like to be able to perform data mining operations. Do you have any plans to introduce internet SQL searching?

dbSNP can not be queried over the internet using SQL. You could try creating your own local dbSNP using the dbSNP FTP files. There is more information about creating local copies of dbSNP in the dbSNP handbook. (**12/18/06**)

Can I query your database server directly?

The NCBI database group policy forbids outside access to dbSNP.

Entrez SNP

Why does the "Common Query Filters" link located on the left blue side bar of the Entrez SNP page lead to the "My NCBI page"?

The link is correct. My NCBI provides a link to commonly used SNP filters, or allows you to set up custom filters for yourself. You will need to set up an account and sign in to "My NCBI" to access the filter resource. **(12/9/05)**

bioMart Type Search Function

Does dbSNP offer bioMart service?

Entrez SNP is the dbSNP version of the bioMart type search function.
Search terms on Entrez SNP can be logically joined using AND, OR and NOT. Here is a simple example:

1. Type "Human[orgn]" (without the quotes) in the textbox at the top of the Entrez SNP page.
2. Click on the "Limits" tab below the text box.
3. Choose the parameters for your search by clicking on the boxes which precede each search parameter.
4. Click the "Go" button located to the right of the text box at the top of the page. (**1/12/06**)

Searching for SNPs in Genes, Chromosomes, Proteins, etc.

Locating SNPs in a Gene or in Genes

Locating SNPs in a Specific Gene

How do I search for polymorphisms in a specific gene?

You could either use Entrez SNP or Entrez Gene as a starting point for your search. To look for SNPs in human epidermal growth factor, for example, using Entrez SNP, do the following:

1. Go to Entrez SNP
2. Type the following search terms in the text box at the top of the page: **[Gene Description] AND [Gene Description] AND [GeneDescription] AND [Organisim]**. Remember to separate each term of the gene name by the Boolean operator “AND”. For example, if the SNPs you are looking for are located in the human Epidermal Growth Factor Receptor gene, your search in Entrez SNP would look like this: **Epidermal AND Growth AND Factor AND human**
3. Click on the first refSNP (rs number) in the result list to get the refSNP page for that rs number, which contains more specific information about that specific refSNP.

To look for a SNP in human epidermal growth factor, for example, using Entrez Gene, do the following:

1. Go to Entrez Gene.
2. Enter the search terms in the text box at the top of the page linked by the Boolean operator AND, followed by the filter: “gene snp”[filter] to retrieve genes that contain SNPs (note: be sure to include the quotation marks in the filter text). Using the human epidermal growth factor example, the search would look like this: epidermal AND growth AND factor AND “homo sapiens” AND “gene SNP”[filter].
3. Click the “Go” button to the right of the text box.
4. On the results page, to the far right of each record is the word “Links” in blue text — click on it to activate a drop-down menu of NCBI resources that are linked to this record.
5. Select “GeneView in dbSNP”, which is located toward the bottom of the menu.
6. You will get the dbSNP report for that locus record.(**04/05/06**)

Where is the resource that allows you to actually see in a graphic how many SNPs are known to lie within a particular gene, and their physical coordinates?

The resource you refer to is the sequence viewer graphical display of a gene region on the human genome. We had to discontinue this resource because the NCBI sequence viewer itself fails when the sequence under consideration is very large (contig sized).

You can still get a sequence view of the mRNA record using the seqview button on the Entrez SNP display, but this won't show intronic SNPs. There may be other inconsistencies even among coding SNPs since the position of SNPs on mRNAs is taken from a different mapping pipeline then the one used to compute genomic locations.

For now, you can search Entrez SNP using the gene symbol or gene id and click the 'GeneView' icon on any of the displayed rs records:

1. Go to Entrez SNP and enter the search terms “LPL AND Human” (without the quotation marks) into the text box at the top of the page, and click on The “Go” button.
2. You will get a page back that contains 158 results to your query.

You can also use Entrez SNP to get a rough text representation of a SNP position in a gene by entering the gene symbol in the search box at the top of the page. When you get the result page, select “chromosome report” from the “display” drop-down menu and “Chromosome Base Position” from the “sort by” drop-down menu. **(6/3/05)**

How do I obtain a list of all the known variants (SNPs, indels, etc.) within a gene, and within the 10 kb immediately 5′ and immediately 3′ of the same gene?

The instructions below will show you how to query the gene database to get the 5′ and 3′ chromosome positions for the gene and then query dbSNP for 10 kb on either side of those positions:

1. Get the position from the graphic display on Entrez Gene.
2. Query using the gene symbol/name or gene/locus ID.
3. Display the result as graphics.
4. Record the chromosome number, the 5′ gene position, and the 3′ gene position.
5. Subtract 10 K from the 5′ gene position and add 10 K to the 3′ gene position.
6. Query Entrez SNP using chromosome number and the newly calculated gene positions.
7. Example: 8[CHR] AND 19807051:19835042[CHRPOS]

How do I search for all the SNPs located in a particular gene?

You can search using the gene symbol on Entrez SNP and by clicking the graphic L link for a record to get an overview all SNPs in the same gene.

Here's an example URL.

How do I find out if there are microsatellite variations in the VDR gene?

1. Go to Entrez SNP, and place the text “VDR” in the search box near the top of the page.
2. Click on the “limits” tab, which is located directly below the text search box near the top of the page.
3. Scroll down to the “organisms” limits section and choose "homo sapiens"
4. Scroll down further to “the “SNP class” limits section and choose "microsat"
5. Scroll up to the top of the page and push the “Go” button.

(**9/6/06**)

Locating a SNP in a Gene Using a Gene Name

Where can I search SNP using gene names? Can I perform these searches in batch form by submitting a text list?

Entrez SNP will allow you to search dbSNP using gene names. EntrezSNP will not, however, allow you to conduct a batch search using gene names. To search using multiple gene names, you can use the operator OR in your query: for example, “LPL OR BRCA1 OR EF2”.

How do I search for polymorphisms in a specific gene?

You could either use Entrez SNP or Entrez Gene as a starting point for your search. To look for SNPs in human epidermal growth factor, for example, using Entrez SNP, do the following:

4. Go to Entrez SNP
5. Type the following search terms in the text box at the top of the page: **[Gene Description] AND [Gene Description] AND [GeneDescription] AND [Organisim]**. Remember to separate each term of the gene name by the Boolean operator “AND”. For example, if the SNPs you are looking for are located in the human Epidermal Growth Factor Receptor gene, your search in Entrez SNP would look like this: **Epidermal AND Growth AND Factor AND human**
6. Click on the first refSNP (rs number) in the result list to get the refSNP page for that rs number, which contains more specific information about that specific refSNP.

   To look for a SNP in human epidermal growth factor, for example, using Entrez Gene, do the following:

7. Go to Entrez Gene.
8. Enter the search terms in the text box at the top of the page linked by the Boolean operator AND, followed by the filter: “gene snp”[filter] to retrieve genes that contain SNPs (note: be sure to include the quotation marks in the filter text). Using the human epidermal growth factor example, the search would look like this: epidermal AND growth AND factor AND “homo sapiens” AND “gene SNP”[filter].
9. Click the “Go” button to the right of the text box.
10. On the results page, to the far right of each record is the word “Links” in blue text — click on it to activate a drop-down menu of NCBI resources that are linked to this record.
11. Select “GeneView in dbSNP”, which is located toward the bottom of the menu.
12. You will get the dbSNP report for that locus record.(**04/05/06**)

How do I find the genomic coordinates of a gene if I know the gene name, for example, scnn1a?

Search for the gene name on Entrez Gene .

Click on the blue text that says “Links” located on the far to the right of the scnn1a result to activate a drop down menu, and select "GeneView in dbSNP".(**6/7/06**)

Is there a bad link to the SNP records for NRP2 (human neuropilin 2)? When I key RP2 into the Entrez SNP search box, it takes me to NELL2, a completely different gene.

By keying only “RP2” into the search box, you are telling Entrez to search not just the gene name field but all fields that contain “RP2”; so you end up with 1465 records, most of which do not pertain to your gene of interest. If you key in “RP2[Gene Name]” into the Entrez SNP search box, then Entrez SNP will retrieve only those SNP records that contain RP2 in the “Gene Name” field (352 records).

Locating SNPs in a Named Gene

I looked up the enzyme "catalase" on Entrez Gene and found the associated SNP page. It states that there are 240 entries for Homo Sapiens. Are there 240 SNP variations that have currently been identified?

There are 240 unique reference SNPs in dbSNP for the human CAT gene. There may be more identified in the literature that have yet to be submitted to dbSNP. (**5/22/06**)

How do I search for SNPs located in the various cytochrome P450 enzymes?

1. There are many ways to use the NCBI web pages to view SNPs located in the cytochrome P450 enzymes. Below are the steps for one way to do this search for a specific gene — CYP27B1 — in the cytochrome P450 enzyme family.
2. Since you have a list of gene products (the enzymes), and you wish to find SNPs associated with those products, I would suggest starting from Entrez gene.
3. Type "CYP27B1" (without quotes) in the text box at the top of the Entrez Gene page and click the “Go” button. You will get the Entrez Gene entry for the CYP27B1 gene.
4. Located on the far right side of the web display there is text that says “links”. Click on the word “links” to get a dropdown menu of information options. Click on the "GeneView in dbSNP" option. This will take you to a page listing the SNPs in and around the CYP27B1 gene.
5. Repeat steps 2 and 3 above for each gene.(**2/8/06**)

How do I determine the number of SNPs located within the human tubulin2 gene (TUBB2)? The accession number is NP_006079.

If you want to locate all the SNPs in and around the TUBB2 gene, you can do so by using one of the following two methods:

Method 1: Go to Single Nucleotide Polymorphism, enter “TUBB2” into the search bar at the top of the page, and dbSNP will retrieve eight SNPs.

Method 2:Go to the refSNP by Locus page. Here, you can see that five SNPs are in the gene region of TUBB2, one SNP is in the genomic region of TUBB2, and GenBank annotation records show two additional SNPs. To have dbSNP show all eight SNPs, click the Send button. This action will send the list to batch query, which will locate and list all eight rs numbers.

How do I find the nucleotide position information for the Leu432Val SNP in CYP1B1? I found its rs number (rs1056836) in Entrez Gene.

1. Go to the SNP homepage.
2. Locate the SEARCH by IDs section of the page, type rs1056836 in the text search box, and click on the Search button.
3. You'll be taken to the rs report page for rs1056836. Go to the GeneView section.
4. Click on the U03688 link. This will take you to the Entrez nucleotide report for this sequence:.
5. Select GenBank from the drop-down list of display options located at the top of the page, and you'll be given the GenBank record for this sequence. Scroll a third of the way down the page to locate the CDS (coding sequence) annotation. Make a note of the start codon for the sequence. Repeat steps 4 and 5 for NM_000104.

You'll find that the start codon for the CYP1B1 gene is located at position 347 for U03688 and at position 373 for NM_000104: CDS 347..1978 for U03688, CDS 373..2004 for NM_000104.

Translated into English, the above records indicate that the RefSeq (U03688) starts 26 bp upstream (the 5′ UTR) of the GenBank mRNA (NM_000104).

You can use the RefSeq record above to determine the position of SNP rs1056836 relative to the start codon by doing the following: Subtract 1 from the start (codon) position, 347 − 1 = 346. Now, subtract 346 from the position of the SNP, 1640- 346 = 1294 exons.

If you use the GenBank mRNA record to determine the position of SNP rs1056836 relative to the start codon, 373 − 1 = 372, then 1666 − 372 = 1294 exons. Divide the number of exons by 3 to determine the codon position of the SNP: 1294/3 = 431.3. It looks like the SNP is in the first position of the 432 codon.

Of course, if any of these sequences were to contain an insertion or deletion with respect to one another, the final positions would be different. It is therefore best to reference a sequence and position rather than just a position.

Nucleotide positions are enumerated starting at 1 for each record shown on the dbSNP page; mRNA and RefSeq positions skip introns.

How do I get a list of SNPs and their heterozygosity reports in table format for the human SPP2 gene?

1. Search Entrez SNP using the [Gene Name] field.
2. Click on the tab marked “Human”, which is located in the second set of tabs at the top of the page.
3. Just above the “Human” tab you will find a menu of display options. Click on the blue arrow located in the “Display” text box to activate the drop-down menu. Select the “Chromosome Report” option.
4. Once the Chromosome Report option has been selected, you will be taken to the chromosome report format for your search results. If the average heterozygosity is available for your SNPs of interest, they will be displayed under the column entitled “avg het”, which is located toward the right side of the page. (**3/24/06**)

How do I find out if there are microsatellite variations in the VDR gene?

1. Go to Entrez SNP, and place the text “VDR” in the search box near the top of the page.
2. Click on the “limits” tab, which is located directly below the text search box near the top of the page.
3. Scroll down to the “organisms” limits section and choose "homo sapiens"
4. Scroll down further to “the “SNP class” limits section and choose "microsat"
5. Scroll up to the top of the page and push the “Go” button.

(**9/6/06**)

Locating a Specific Variation Type in a Named Gene

How do I find out if there are microsatellite variations in the VDR gene?

1. Go to Entrez SNP, and place the text “VDR” in the search box near the top of the page.
2. Click on the “limits” tab, which is located directly below the text search box near the top of the page.
3. Scroll down to the “organisms” limits section and choose "homo sapiens"
4. Scroll down further to “the “SNP class” limits section and choose "microsat"
5. Scroll up to the top of the page and push the “Go” button.

(**9/6/06**)

Locating SNPs in Specific Regions within a Gene

How do I search for SNPs located in a specific region of interest within a gene?

There are many different ways to focus on SNPs located in a specific genomic region. Here is an example, using the human ABO blood group locus as an example.

From the Entrez SNP webpage, do the following:

1. Type “ABO[Gene Name]” (don't include the quote marks) in the for textbox located in the gray area at the top of page.
2. Click Go. The default display is a list of SNPs with a graphical display. In this case, you should get a list of 241 SNPs that dbSNP is able to map to the ABO locus (which should include the promoter region).
3. Use the Select box located above the list to choose one of several display formats. Once you have selected a format, click the Display button. Selecting the dbSNP Batch Report display option will allow you to access several downloadable formats.

You can access the dbSNP GeneView page by clicking on the blue L link shown with any of the displayed SNPs.

How do I find the sequence position of a SNP on a gene?

As far as I know NCBI doesn't have a canonical set of "gene records" that would provide a coordinate system for SNP alignment to genes. NG_ records are curated gene regions, but they do not exist for all genes annotated on the human genome. In any event, we make no systematic effort to map our SNPs to this set apart from NCBI genome alignment.

dbSNP maintains a table called ContigExon which provides the start and stop positions of each exon in contig coordinates for each mRNA refseq defined in a particular NCBI genome build. We use these coordinates to create the exon placement graphic on our Geneview display.

You can obtain the same canonical list of exon boundary information in chromosome coordinates from an NCBI genomes ascii file [ftp.ncbi.nih.gov/genomes/H_sapiens/maps/mapview/BUILD.35.1/seq_gene.md.gz]. **(4/20/05)**

Locating SNPs in Genes from a Specific Population

How do I search for known SNPs in several genes that occur only in the European/Caucasian population?

Although dbSNP does not have a classification for race and ethnic group, you can search on Entrez SNP for the gene and limit the subset to population class EUROPE. Enter the gene name or term in the search box, click Limits, and check the box for EUROPE under limit by Population Class.

Locating Genes that Contain a Large Number of SNPs

Is there any way of identifying genes that contain a large number of SNPs or retrieving SNP data by using the number of SNPs/gene as a query term?

dbSNP doesn't have a web service for this kind of search. You will have to download the SNPContigLocusId table [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/b125_SNPContigLoc_34_3.bcp.gz], located in the organism_data directory for your organim (human in this case), and query it using SQL. **(4/20/05)**

Locating SNP Information for a Large Number of Genes

How do I get SNP information (refSNP number, contexts, SNP type, etc.) for 1000 genes?

You'll have to write a program using eutils programming utilities:

1. Use eSearch to perform the search and parse out the refSNP number. Look at the following example, which shows a search for SNPs in the human LPL gene (txid9606).

2. Use eFetch to retrieve SNP report with the refSNP number from step 1. Look at the following example, which shows the retrieved fasta report for rs17410783.

1. You'll have to put your query after the term parameter and remove the LPL gene symbol, since I included it only as an example. **(2005)**

Analysis of SNP Data Associated with a Specific Disease

Could you suggest a statistical program to analyze data from a SNP that may be associated with a certain kind of heart failure?

There are many different statistical packages that could perform the type of analysis that you need to do. A comprehensive list of all types of genetic software is available.

Rather than suggesting a particular statistical package for the evaluation of allelic association with a complex phenotype such as heart disease, I would suggest that you involve a statistician for the actual analysis because there are pitfalls to conducting such research without proper statistical knowledge. Please take a look at the following review: *The Lancet: Problems of reporting genetic associations with complex outcomes* [PubMed].

I searched dbSNP for "diabetes mellitus" and got 2585 SNPs. How do I determine the number of genes on which these SNPs are located?

Go to the Locus Information Query page. Choose human from the organism list and type in "diabetes mellitus" in the query box. The search will take you to LocusLink, which will show a count of 66 loci; this is the number of genes relating to "diabetes mellitus" that have SNPs.

The 2585 SNPs in your search result are the number of hits to the LocusLink description "diabetes mellitus". Very few dbSNP records have phenotype and disease descriptions. That is why you don't see "diabetes mellitus" mentioned in dbSNP records of these SNPs.

Locating SNPs Associated with Disease Causing Mutations

I want to locate SNPs in dbSNP that have been associated with familial dilated cardiomyopathy. Is it possible to narrow the search on dbSNP to mutations that have been published in the literature as disease-causing mutations?

dbSNP does not currently support the query of MUTATIONS or other variations with relationships to disease. Below, I will provide an explanation of our current data and offer some ideas on capturing SNPs in genes if a focus on a limited number of genes is justified by your disease model.

The great majority of data in dbSNP has been collected and defined as variations simply using sets of co-aligned genomic or DNA sequences. Because this process typically has little to no focus on disease condition, only about 250 records in dbSNP have been successfully associated with phenotype-causing mutations or a clinical outcome in OMIM. SNPs are richly connected with genes on human build 34, however, and this connection may confer "candidate status" to otherwise anonymous submissions as noted below.

I would expect that you will have better success in searching disease-oriented resources, such as OMIM and HGMD, for the topic of familial dilated cardiomyopathy. Both resources curate lists of published allelic variants or "mutations" (typically defined at the protein level with amino acid substitutions for "alleles" instead of nucleotides as in dbSNP) that have been associated with genes and/or phenotypes. Note that OMIM restricts its curation to "important" allelic variants, whereas HGMD attempts broader coverage of mutations reported in the literature. Additionally, the Human Genome Variation Society (HGVS) maintains a web-based collection of locus-specific mutation databases related to particular genes or diseases.

It is important to recall that many if not most mutations reported in the literature have been detected in simple Mendelian disorders. If familial dilated cardiomyopathy is more properly considered a complex genetic disease involving a large number of genes with small individual effects, then the likelihood of finding large numbers of "relevant" or "significant" mutations in any of these resources is necessarily diminished. In the event you elect to research a complex trait, I propose the following thoughts as a potential framework for moving from published mutations into SNPs and Haplotype data as the need arises:

First and foremost, you should review your disease model (e.g., the number of genes with major and minor effects, penetrance of phenotype, complexity of gene × gene and gene × environment interactions, etc.) and determine how many genes from the model would have identifiable alleles of major effect. The number of genes on this list can measure the suitability (if small) of adopting a molecular epidemiology candidate gene approach, where inferences about diseases are made through computations and assessments of the relative risk in having one allele instead of another. This approach combines statistical and molecular constraints on the data to infer functional pathways, mechanisms of causality, and the "risks" of disease for each polymorphism in gene sequence (discussed below).

If you accept the premise that the magnitude (if large) and number of "major" allelic effects specified by an accurate disease model should exhibit agreement in number and strength of effect with the set of mutations recovered from resources such as OMIM and HGMD, then an early assessment of allelic effects could:

- provide a baseline expectation or “estimated yield” prior to scanning mutations from public resources
- demonstrate when a set of recovered records is smaller than predicted by the disease model and hence potentially incomplete
- demonstrate that the set of recovered records is sufficiently dense for research needs
- demonstrate that the model is incommensurate with data and must be revised

For example, if the analysis of your disease model indicates that a limited number of genes are involved and that there are specific, detectable alleles with strong effects, I would propose that additional searches of PubMed may reveal lists of genes known or hypothesized to be involved in your cardiomyopathy phenotype.

If a candidate set of genes can be defined, then you can, in turn, query dbSNP for the set of reported polymorphisms spanned by the appropriate mRNA feature. Keep in mind that although this later set may have little or no direct evidence for causality in any phenotype (e.g., the SNPs were not ascertained in probands; no publications currently reporting them as mutations), these variations would be a rational choice for follow-up *in silico* assessments of their disease-causing potential. The subset that survives this final cut could be adopted for validation in a real clinical context.

We have moved some of the allelic variant content of OMIM into dbSNP, so the sequence is available, and the variant position can be confirmed by sequence alignment. I also understand that some groups at EBI have undertaken projects to electronically mine SNP–gene pair relationships directly from titles and abstracts in MEDLINE. This might be foreshadowing another class of tools/resources that captures the SNP–gene–disease relationships described in the full body text of journal articles.

We are trying to find data on drug resistance-causing mutations. Are you aware of a database that lists submissions by protein, organism, point mutation, and drug?

We are not aware of a database that lists specifically the data you requested, but you could try searching OMIM for mutations affecting drug resistance.

Although the data you require are not presented strictly as a list, you may have to read the original publication listed on OMIM to get the details. You could also try searching Google for “drug resistance mutation database”. Google search

Locating SNPs in Promoters

How do I locate SNPs in promoter regions?

SNPs are divided into the functional classes, shown below. This list does not include promoter regions. We'll look into adding promoter regions in the future.

SNP functional classes:

- locus region
- intron
- coding nonsynon
- mrna utr
- coding synon
- reference
- exception
- splice site

How do I find the SNP location for the G/C polymorphism in the IL6 promoter located at base 174 and the G/C polymorphism found in the following sequence: ttgtgtcttgc(G/C)atgctaaa ggacgtcaca ttgcacaatc ttaataaggt ttccaatcag?

SNP locations are stored as chromosome positions in dbSNP. Use the sequences flanking the variation of interest to MegaBlast against the chromosome database; the results will include the chromosome number and the position where the SNP of interest is located. Once you have the chromosome number and the position, you can use them to query Entrez SNP. Please see the example below:

Example:

1. Use Megablast to get the chromosome position(s).
2. Use the chromosome number and position(s) to query Entrez SNP. This link shows “no items found”. This means that the SNP has not yet been submitted to dbSNP.
3. If the variation had been submitted to dbSNP, you would have found a “submitted SNP” accession number (ss#), or “reference SNP” number (rs#) cited in the literature. Here is a link to other SNPs located in IL6 that you might be interested in.

How do I find the starting and ending positions of a gene?

To obtain gene starting and ending positions on a contig, follow the instructions below once you get your SNP search results:

1. Select Display, then Gene Links on the Entrez SNP page (toward the top).
2. Select the gene of interest on the Entrez Gene.
3. Select Display, then Gene Table on the Entrez Gene page (toward the top).
4. Make a note of the starting and ending positions on the graphic and use them to search on Entrez SNP using CTPOS.

If you know the position of your regulatory region, then search with the contig position [CTPOS] using the upstream offset from the starting position of your gene. You can also use the chromosome position [CHRPOS] in conjunction with the chromosome number field [CHR].

Locating SNPs in Exons/Introns

After locating SNPs within a gene, is there a way to scan for SNPs in exons vs. introns, and missense vs. silent polymorphisms, without examining each SNP record?

Yes, use Entrez SNP. Below is an example of this quick scanning technique using the CLOCK gene in humans:

1. Go to the Entrez SNP search page or the dbSNP homepage.
2. Enter CLOCK[GENE] AND HUMAN[ORGN] in the Search box and click Go.
3. 99 SNPs will be returned in the search result. There is a graphic bar under each snp_id in the search result. Within that bar, there will always be an L, which stands for “locus”. If the L is blue, then the SNP is in the locus (our search only returned SNPs at a gene locus).
4. Just to the right of the L is a T, which stands for “transcript”. If the T is blue, then the SNP is also in the transcript. To the right of the T is a C, which stands for “coding region”. If the C is any color other than wash-out white, then the SNP is also in a coding region. If the C is green, then the SNP will produce a non-synonymous change in the protein. If the C is red, then it is a synonymous change.
5. You can restrict a query to a particular type by checking the appropriate boxes on the Limits form found in the function class section of the page. Access the Limits form from the Limits link, directly under the search form.
6. You can ask for the union of those SNPs that either code synonymous or are located in introns to be returned by checking the corresponding boxes. There are 42 such SNPs for the CLOCK gene. Or, you might only be interested in non-synonymous changes for the CLOCK gene; in this case, set the coding non-synonymous box on the Limits form, and then you’ll find that only two SNPs are returned, rs3762836 and rs1056478.

How do I search for verified SNPs that are contained on expressed material (5′ UTR, 3′ UTR, or exons) and are polymorphic between B6 and BALB/c PLUS mice?

Go to the Search Mouse SNP between strains tool and limit the results on Entrez using the function class and genotype filter.

How do I find exon starting and ending positions within a particular gene?

We have a file with the exon starting and ending positions within a contig in table ContigExon. You can get the file ContigExon.bcp.gz in the organism_data [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data] (link goes to human_9606 database as example) directory found within your organism’s database.

How do I determine if there are any SNPs in the first exon/intron in the factor ix human gene?

Please follow the search instructions below.

1. Search Entrez SNP using the terms "factor ix AND human".
2. Click on the L graphic link for the first SNP on the page; this will take you to the summary page for SNP → LocusLink.

How do I find validated SNPs located in coding sequence that have heterozygosity over a certain threshold, say 30%?

1. Make a list of gene IDs (the gene ID for the example below is 4023) and look them up on Entrez Gene.
2. Upload the list of gene IDs on this page. Click retrieve to show the gene results.
3. Select SNP links and click the Display button to see the results.
4. Click on Limits at located at the top of the page, and select your filters (i.e., validated, coding, or heterozygosity).

How do I search for information regarding the number of SNPs found in coding regions within a gene using the gi, NM, or XM ID numbers?

Go to the organism_data [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data] directory located in your organism's database. FTP Data page [ftp.ncbi.nih.gov/snp/mssql/data] and click on SNPContigLocusId.bcp.gz. This is an ASCII image of the database table containing all of our snp-to-gene information as gathered from the NCBI assembly processing. See below for column headings.

You probably only want to keep rows where fxn_class is 3, 4, or 8.

Here are the columns. The example is the first NON_SYNON encountered from the top:

Columns	Example	COMMENT
snp_id	47	snp_id
contig_acc	NT_007819	contig accession
contig_ver	14	contig version
asn_from	10876794	contig position
asn_to	10876794	contig position
locus_id	23249	locus link identifier
locus_symbol	KIAA0960	gene name
mrna_acc	XM_371877	mrna accession
mrna_ver	2	mrna version
protein_acc	XP_371877	protein accession
protein_ver	2	protein version
fxn_class	4	2,3,4,8 are coding, 6=intron, 5=utr
reading_frame	1	
allele	G	allele as found on mrna_acc
residue	D	residue as found on protein_acc
aa_position	514	position of residue I protein_acc
build_id	34_3	Genome build; assembly context is a property of contig

Locating SNPs in Stop Codons

How do I format a search for all the stop codons in dbSNP?

I do not know of a way to query either the dbSNP homepage or Entrez SNP to get that information. If you have a local copy of the dbSNP database, you can do the following query:

Select "snp_id,locus_symbol, protein_acc FROM SNPContigLocusId, WHERE residue = '*'" to get all the SNP– protein relationships (that we can identify) for organism=human, where the SNP is in the stop codon. If you don't have a local copy of the dbSNP database, you can parse the same information from the SNPContigLocusId table located in your organism's organism_data [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data] (this link is to the human_9606 directory as an example). (see columns 1,7,10).

How do I find the starting and ending positions of a gene?

To obtain gene starting and ending positions on a contig, follow the instructions below once you get your SNP search results:

5. Select Display, then Gene Links on the Entrez SNP page (toward the top).
6. Select the gene of interest on the Entrez Gene.
7. Select Display, then Gene Table on the Entrez Gene page (toward the top).
8. Make a note of the starting and ending positions on the graphic and use them to search on Entrez SNP using CTPOS.

If you know the position of your regulatory region, then search with the contig position [CTPOS] using the upstream offset from the starting position of your gene. You can also use the chromosome position [CHRPOS] in conjunction with the chromosome number field [CHR].

Locating SNPs in Intergenic Regions

How do I obtain SNPs located in human intergenic regions?

Search using the limit strategy below:

1. Enter “all[sb]” to select all SNPs.
2. Click on Limits, then check and enter the following limits: CBID Range from 1 to 119, chromosome 1, Homo sapiens, snp[SnpClass], Weight 1. Result for query #1 = 393170.
3. Click on Limits again, then check and enter the following limits: CBID Range from 1 to 119, chromosome 1, coding nonsynon, reference, exception, intron, coding synonymous, locus region, mrna utr, splice site, Homo sapiens, snp[SnpClass], Weight 1. Result for query #2 = 180392.
4. Click on History, which shows you the queries and results.
5. Uncheck the Limits box.
6. Use the search numbers located next to each query to do a Boolean query. For this example, assume your queries above have been assigned the search numbers “#1” and “#2”. Enter in the search box “#1 NOT #2” without the quotes. Result for query #3 = 212778.

Locating SNPs in (a) Chromosome(s)

I need to identify all human Y chromosome specific SNPs. What filter do I use to remove SNPs that also map to the X chromosome?

Add the filter “Weight 1” to retrieve SNPs that only map to the Y chromosome.(**11/27/06**)

Where can I access a map of a specific chromosome region and use it to search for SNPs within that region?

You can access chromosome maps by going to the NCBI home page, and clicking on the Mapviewer link located in the alphabetic list of links located on the right hand side of the page.

1. Once you have reached the Mapviewer page, click on the button labled "Switch to graphical view" located near the top of the page.
2. Once you have the graphical representation, click on an group of interest to you to see a dropdown menu of the organism maps available. Select an organism. For this example, I selected “Mammals” and then “Homo Sapiens Build 36” to see a graphical representation of the organisim's (Homo Sapiens) chromosomes.
3. Select a chromosome by clicking on the chormosome number located below the chromosome graphic. This will give you a view of the chromosome that includes an ideogram, a contig map, a Unigene map and a gene sequence map. You can alter the view by clicking on the "Maps and Options" button located on the upper right hand corner of the page.
4. You can search specific regions of the chromosome you selected by clicking on the section of the chromosome ideogram you want to view more closely, and then select the magnification factor you would like to view the section under from the resulting dropdown menu.(**11/27/06**)

Does dbSNP have a single flat file available for download that contains the chromosome assignment(s) and physical positions(s) for each rs entry?

Yes. There is a flat file called the “Chromosome Report”, located in each the dbSNP FTP site for each organisim. “Chromosome Report” provides an ordered list of RefSNPs in approximate chromosome coordinates (the same coordinate system used for the NCBI genome MapViewer). The column definitions for “Chromosome Report” are located in the dbSNP FTP readme [ftp.ncbi.nih.gov/snp/00readme.txt] file. Once in the readme file, scroll down until you come to the “Chromosome Reports” section. **(12/01/05)**

Which dbSNP report format contains the location and other important information for each SNP?

The “chromosome reports” (chr-rpts) report format located in the dbSNP FTP site reports the map locations of all SNPs in both contig and chromosome coordinates. You can access this file using the following steps:

1. Go to the dbSNP FTP site [ftp.ncbi.nlm.nih.gov/snp], and click on your organism of interest. You will be taken to an index of the reports available for that organism. Please note that not every organism has data available in the “chromosome reports” report format.

2. Click on chr_rpts. This will take you to a list of the available chromosome reports, listed by chromosome, for the organism of interest. Please note that not every chromosome for a particular organism may have a chromosome report.

3. Documentation for the columns found in the “chromosome reports” report format can be found in the general read-me file [ftp.ncbi.nlm.nih.gov/snp/00readme.txt], under the heading “CHROMOSOME REPORTS”.

How do I download a list of all SNPs (confirmed and unconfirmed) in a region between positions 151983001 and 152252001 on chromosome 1?

1. Go to the Entrez SNP page.

2. Enter one or more search terms in the text box at the top of the page.

3. Click on the “limits” tab located just below the text box to restrict your search by search field, chromosome, and other criteria. Please note that the field descriptions and class definitions you will need to know in order to formulate the correct search restrictions are located on the Entrez SNP page — just scroll down below the big SNP bubble graphic and you’ll find them.

4. Go to the “Chromosome(s)” section of the limits page (second section on the left), and enter 151983001in the “Base position from” text box, and then enter 152252001 in the "Base Position to” text box.(**3/13/06**)

How do I find mouse SNPs that are associated with specific regions on different mouse chromosomes?

Here is an example that shows how to do this kind of search:

1. Start on the Entrez SNP home page

2. Type "Mouse[orgn]" (don't include the quotes) in the text box at the top of the page.

3. Now, click on the “Limits” tab --- doing this will generate a form that contains many selection options

4. Go to the “Chromosome(s)” section of the Limits page and click the “chromosome 1” check box.

5. At the bottom of the“Chromosome(s)” section of the Limits page and key in "1" (don’t include the quotes) in the “Base Position: from” text box, and "100000"(don’t include the quotes) in the “Base Position: to” text box.

6. Go to the “Map weight” section of the Limits page and click the number “1”check box (checking this box means that the SNPs will map uniquely)

7. Go to the top of the page and click the “Go” button. You should get back a list of 141 SNPs for this particular search.

8. You can now select the display (report) format by clicking the “Display” drop-down menu toggle located just underneath the tabs on the left near the top of the page, and clicking the format of your choice. **(4/22/05)**

How do I find all the SNPs located within a particular chromosomal position?

You can get the SNPs for a particular chromosomal position by using Entrez SNP and by specifying the chromosome and position(s). Here’s an example.

You can also use the Limit function to filter your result by chromosome and position. When you receive a search result, you can view the chromosome report for that result by changing the display option on the report.

I am having difficulty searching for SNPs between 88,000,000 and 93,000,000 bp on mouse chromosome 6. Can you help me?

Key in the following search in the search text box at the top of the Entrez SNP page: "Mus musculus"[ORGN] AND 88000000:93000000[Base Position] AND 6[CHR] **(5/5/05).**

How do I search dbSNP for the total number of SNPs located between nucleotide positions “x” and “y” on the short arm of human chromosome 18?

Below is the suggested search for your question.

1. Search Entrez SNP for the term “human”.
2. Click Limits.
3. Select chromosome 18 and enter the nucleotide positions “x” and “y” in BasePosition from and to, respectively.
4. Click Go.

How do I update the chromosome positions of microsatellite primers?

You can get the chromosome positions from the “Chromosome Report” using Batch Query.

How do I use dbSNP refSNP (rs) numbers to get a list of chromosome positions?

1. Go to the dbSNP home page, and scroll down to the “Batch” section.
2. In the “upload list” sub-section, select “Reference SNP ID (rs)”.
3. On the data request form, enter the email address where you want your data sent, select the organism of interest from the “organism” dropdown menu, and in the “Select Result Format” section choose “CHROMOSOME RPT” from the dropdown menu (3/1/06).

How do I find SNPs located on chromosome 7 between markers DXMIT171 and XMIT235 in C3H and Black 6 mice?

Find the location of markers DXMIT171and DXMIT235 using uniSTS.
Markers DXMIT171 and DXMIT235 are located on chromosome X at 8769971 and 141938910, respectively. Use the SNP homepage search form and do the following:

1. Type in “X[chr] AND 88769971:141938910 [Base Position]” in the text box located at the top of the page.
2. Click on limits, which is located below the textbox, to select your options. In this case, select *Mus musculus* from Organism, and specify that weight = 1 (uniquely mapped).
3. Press Go.
1. The SNP homepage search will produce a list of genotyped SNPs between your markers that are mapped to the mouse genome.
4. Now click on the G link for each SNP to view the genotype report.

Please note that C3H/HEJ mice are not always typed.
You can also use the XML-based mouse genotype reports [ftp.ncbi.nih.gov/snp/organisms/mouse_10090/genotype] to find differences between various strains. Please read the readme00.txt for information on the XML format.

Locating SNPs in a Protein Family

How do I search for a particular protein family that contains SNPs?

You can search for proteins that have SNPs from the NCBI protein database using the protein_snp subset.
Example:

1. Search Entrez Protein using the protein accession and protein_snp subset (sb) as search terms.
2. Then search Entrez Protein using the protein found in the previous step and protein_snp[sb] as search terms.

3. Select Links located on the far right side of the page for the protein of interest, and when the drop-down menu appears, select SNP. This will take you to a SNP page that contains a list of the SNPs for the protein entry.

4. Locate the display box at the bottom left-hand corner of the page and choose one of the SNP reports (FASTA, ASN.1, Flat File, or XML), which will easily allow you to parse the data.

Searching for SNP Primer Sequences

Can I find primer sequence in dbSNP that will allow me to amplify a SNP locus?

dbSNP doesn't provide the sequence of primers. Each SNP submission, however, contains flanking sequence (FASTA sequence usually between 100-500bp) that could be used to design primers. Here is an example of such flanking sequence. **(12/16/05)**

How do I use dbSNP to determine what primers were used for each mouse SNP?

dbSNP does not contain SNP primer information. I recommend that you contact the submitter for this information. We do have plans to capture this information in the future.**(2/3/05)**

I couldn't find PCR primer information for SNPs in dbSNP using the submission report links or LinkOuts in the submission reports, so how do I find them?

When a dbSNP submitter submits an STS primer or provides an STS accession, we link to dbSTS to show the primer information, for example, ss3.

Click on view Detail in dbSTS.

About 300,000 of the submitted SNPs (of a total 8.6 million) have STS information.

You can also search for primers by using the Map Viewer with both the variation and the STS tracks turned on. First, locate the SNP of interest, then use the adjacent STS map to find the STS, and then get the STS accession and primer information by using the uniSTS page.

Here is an example using rs2665.

Go to Map Viewer (click on chromosome 9), find the variation, and then click on STS D9S1131, which takes you to the uniSTS page. Clicking on the GenBank accession number will take you to the GenBank record, where you will find the primer information.

How do I find the rs numbers for two primer sequences? The primers are not attaching very well, and I'd like to extend them.

Try BLASTing the flanking (primer) sequences against dbSNP to see if there are any matching rs numbers. If there are, use the flanking sequence included in the SNP record of the rs number that matches your primer. **(11/15/05)**

I have been searching dbSNP using a SNP accession number (2979099) to find information on primers and assay conditions. I can't find them. Can you help me?

Go to the dbSNP homepage and select Search by IDs, then enter the prefix "ss" next to the accession number (2979099) in the search box. You will then be taken to the Submitted SNP(ss) Details page for that accession number. To view the details of the experiment, locate the ASSAY section, and click on the link that follows the method "PROTOCOL_1".

How do I update the chromosome positions of microsatellite primers?

You can get the chromosome positions from the "Chromosome Report" using Batch Query.

Searching for SNPs or Specific SNP Data Using a Specific ID

Locating SNPs using Celera IDs

How do I use Celera ID numbers to find dbSNP refSNP (rs) numbers?

1. Go to the SNP home page
2. In the “Search by IDs” section of the home page, type the Celera ID number in the text box, and select “Celera SNP ID” from the drop down menu to the right of the text box.
3. Click on the “Search” button.(**3/3/06**)

How do I search for rs numbers that correspond to a list of hCV (Celera) SNP numbers? When I tried to “Search by ID", I got an ss number but not an rs number.

The reason why your list of hCV numbers has associated submitted SNP (ss) numbers but not refSNP (rs) numbers is that these particular variations have not yet been mapped and clustered.

rs numbers have not been assigned for these ss numbers in the current build. They will be automatically assigned during the next dbSNP build mapping and clustering process, and will be available when the new build is released. Information about the dbSNP rs numbering and clustering process is available in the dbSNP handbook.

To receive notification of new build releases, subscribe to dbSNP-announce. (**2/16/06**).

Locating SNPs using STS IDs

What do I enter in the "Between Markers STS Search" field in the dbSNP search window?

Enter the name or ID numbers of an STS (i.e. SHGC-156091 and D1S3135) in dbSTS. **(1/7/05)**

How do I find the refSNP (rs) numbers for Y chromosome markers published by the Y-chromosome consortium (e.g. what is the corresponding rs number for STS DYS190)?

1. Enter "DYS190" into Entrez cross-database search, and click the “Go” button.
2. You will find that there is a single hit (AF337053), which is located in the NCBI Nucleotide database. Click on the Nucleotide Icon to go to the Nucleotide result page.
3. Click on the AF337053 link to get to the Sequence View page for this accession number. You will find the relationship between DYS190 and the human Y chromosome in the “features” section of this page:/ organism="Homo sapiens"/mol_type="genomic DNA"/db_xref="taxon: 9606"/chromosome="Y"/ clone="3-11; DYS190"
4. Click on the word“Links” at the top right corner of the sequence view page; clicking on it will release a dropdown menu. Select “Mapviewer” to see that AF337053 is located on the q11 region of Y.
5. Click on the Y chromosome link to go to Mapviewer. Click on the “Maps & Options” button to add the "Variation" option to your Mapviewer display. Two SNPs are within 2000 bp of AF337053.(**2/14/06**)

Every time I search for STS markers in dbSNP, dbSNP returns a screen that says that no SNPs were found in my region of interest. What am I doing wrong?

I tested the STS search and it works for the query: between STS “RH98513” and “SHGC-74072”. For your query to work, the STS markers you use must exist in dbSTS and must be on the same chromosome. Check your markers against dbSTS to make sure that they are valid.

Locating SNPs using PubMed ID

If I know the PubMed number of a SNP, how do I get the rs ID of this SNP?

Query Pubmed using the ID (pmid). Once you get the result, you can either select SNP Link from Display, or you can or click on the blue word Links at the top right side of the result and select SNP, if it is shown. The SNP option is not shown if there is no SNP associated with the publication.

Here's an example of a publication with a SNP association.

Locating SNPs using CGAP IDs

How do I search for a SNP that is associated with a CGAP identifier?

dbSNP does not contain CGAP identifiers for one of two reasons: either SNPs with CGAP identifiers have not been submitted to dbSNP, or the SNPs were submitted using a different identifier.

If you know the identifier that a particular submitter has used in their submission to SNP, you can search for SNPs of interest using this identifier:

1. Go to the dbSNP homepage
2. Go to the Search by IDs section, and enter the SNP indentifier in the text box.
3. Select "Submitter SNP ID" from the drop-down menu to the right of the text box and click on the "search" button located below the text box.

If you have the sequence, you could also try a BLAST search. (**3/14/06**)

Locating Specific Data Using RefSNP(rs) or Submitted SNP(ss) ID

If I have a refSNP ID number, how do I find the trivial name used by the submitter in published manuscripts for that variation?

Use the "Search by IDs" query module that is located on the dbSNP homepage. It will allow you to select SNPs based on dbSNP record identifiers, including: the reference SNP (refSNP) cluster ID numbers (rs#), submitted SNP Accession numbers (ss#), and Submitter SNP (trivial) IDs for the same variations.

To use Search by IDs:

1. Enter the ID in the input text box and select the ID type from the drop down menu located next to the input text box. Now click "Search".
2. The result page will show the reference SNP (refSNP) cluster ID numbers (rs#), submitted SNP Accession numbers (ss#), and Submitter SNP (trivial) ID for the ID you submitted. **(11/17/05)**

How do I use dbSNP refSNP (rs) numbers to get a list of chromosome positions?

4. Go to the dbSNP home page, and scroll down to the "Batch" section.
5. In the "upload list" sub-section, select "Reference SNP ID (rs)".
6. On the data request form, enter the email address where you want your data sent, select the organism of interest from the "organism" dropdown menu, and in the "Select Result Format" section choose "CHROMOSOME RPT" from the dropdown menu (3/1/06).

How do I use a list of gene ID numbers to get dbSNP information in an automated fashion, similar to using Efetch Eutilities to query Entrez?

You can use the [LID] field in the eSearch query below. The [LID] field is synonymous with GeneID.

Example:

Search for SNPs in the gene LPL with gene/locusID=4023.

You can then use the returned QueryKey and WebEnv numbers in Efetch to retrieve the reports.

I want to retrieve GeneBank accession numbers associated with a list of rs ID and ssID numbers. Do you have a SNP text file containing this information that I can parse?

There is an eLink utility that will allow you to retrieve GenBank records associated with an rs number. Please see the online instructions and examples for Elink. **(10/27/05)**

I entered dbSNP rs numbers for a region and got back 129 entries. When I looked a few of them up, they were returned as "not found", although I know there are data for them.

Please try entering only the number and the "uid" field without the "rs" prefix (i.e., 10168104[uid]).

I have been searching dbSNP using a SNP accession number (2979099) to find information on primers and assay conditions. I can't find them. Can you help me?

Go to the dbSNP homepage and select Search by IDs, then enter the prefix "ss" next to the accession number (2979099) in the search box. You will then be taken to the Submitted SNP(ss) Details page for that accession number. To view the details of the experiment, locate the ASSAY section, and click on the link that follows the method "PROTOCOL_1".

How do I find the "1066A>G" notation for a SNP whose dbSNP id is known (e.g. rs12695902)?

dbSNP does not use the "1066A>G" notation or offset from the start codon (ATG) to describe a SNP position, unless it is given to us by the submitter. You can use the "GeneView" or "SeqView" display options for a refSNP (rs) number to see if the SNP is in available in either the gene context, or is available in a coding region position:

1. Search Entrez SNP for your rs number:
2. Click on either "GeneView" or "SeqView" in the small graphics located below the refSNP (rs) sequence. (**3/14/06**)

Locating SNPs using IDs Published in the Literature

How do I find a refSNP (rs) number for a SNP using a reference number found in the literature?

The dbSNP GeneView (SNP linked to Gene) page may be the best way to look for a refSNP (rs) number rs that corresponds to published gene position variations. You can reach the GeneView page by clicking on the pink "GeneView" button located under the refSNP (rs) number of interest in the results page of a Entrez SNP search. Using the GeneView page may not always work if the published gene model is different that the model used in dbSNP, but in your case there is strong evidence that the SNP in position 45 is rs2241766, which is shown on the GeneView page as amino acid 15, codon position 3 (3*15) = 45. rs2441766 is located at contig position 93066042.

Looking for the other two variations is a little more difficult. First, click on the radio button labeled "in gene region", which is located in the long, grey, rectangular box found above the gene model (contig mRNA transcript) section of the GeneView page. Now click on the "refresh" button located at the far right end of the grey box where you found the in gene region" button. You will now see all SNPs near ADIPOQ (the default is cSNPs). The SNP in position 713 is likely rs3774261. I determined this by subtracting 45 from 713 (the difference in the position of the two variations) and then adding this number to the contig position of the SNP at position 45. (713-45)+ 9306642 = 93066709. rs3774261 is located at contig position 93066709. (**6/1/06**)

How do I find a refSNP (rs) number for a SNP using a reference number found in the literature?

The dbSNP GeneView (SNP linked to Gene) page may be the best way to look for a refSNP (rs) number rs that corresponds to published gene position variations. You can reach the GeneView page by clicking on the pink "GeneView" button located under the refSNP (rs) number of interest in the results page of a Entrez SNP search. Using the GeneView page may not always work if the published gene model is different that the model used in dbSNP, but in your case there is strong evidence that the SNP in position 45 is rs2241766, which is shown on the GeneView page as amino acid 15, codon position 3 (3*15) = 45. rs2441766 is located at contig position 93066042.

Looking for the other two variations is a little more difficult. First, click on the radio button labeled "in gene region", which is located in the long, grey, rectangular box found above the gene model (contig mRNA transcript) section of the GeneView page. Now click on the "refresh" button located at the far right end of the grey box where you found the in gene region" button. You will now see all SNPs near ADIPOQ (the default is cSNPs). The SNP in position 713 is likely rs3774261. I determined this by subtracting 45 from 713 (the difference in the position of the two variations) and then adding this number to the contig position of the SNP at position 45. (713-45)+ 9306642 = 93066709. rs3774261 is located at contig position 93066709. (**6/1/06**)

Using SNP IDs to Find Genbank Accessions, and Vice Versa

Is SNP information annotated in GenBank entries? If so, how do I access this annotation?

dbSNP annotates GenBank reference sequences and mgc mRNAs based on BLAST analysis, rather than on submitter accessions. Reference Sequences, or RefSeqs (rs), are a curated, non-redundant set of records for mRNAs, proteins, contigs, and gene regions constructed from a GenBank exemplar for that protein or sequence.

I will use rs11542065, a non-synonymous variation in the LPL gene, as an example. If you look under GeneView, you'll see that the SNP maps to protein NP_000228. If you click the link for this protein, it will take you to the GenBank protein record. From here, click the features button at the top of the page, and then check SNP in the checkbox. Once this is done, scroll down, and find the variation blurb for this record:

variation 71/replace="h"/replace="q"/db_xref="dbSNP:11542065"

If you repeat this exercise for the mRNA NM_000237, you will find the variation annotation:

Variation complement(387)/gene="LPL"/replace="g"/replace="c"/db_xref="dbSNP:11542065"

Always remember to set the SNP feature checkbox whenever you load a GenBank record.

Does dbSNP have a search that will take a sequence accession (or gi number) as input and will return the SNP information for that sequence?

I'm sorry that we don't have the exact service you requested. We try to provide resources (search service and reports) that are broad and generic enough to suit a broad range of users. The information that you requested is available, but it will require parsing a dbSNP XML or ASN report that you can retrieve automatically using Entrez Eutils.

Please read the online document and examples for automatic search/retrieval using the Eutils web API.

Here's an example of using eLink to get all the SNP using the protein LPL gi(4557727).

Here's an example of retrieving an XML report for a snp id (rs).

Additional Entrez Tools are available.

I see in "Submitter-Referenced Accessions" that Genbank accession numbers are linked to the submitted SNP (ss) ID number, but I can't find these Genbank accessions in the XML formatted reports.

The accession numbers you see under "Submitter-Referenced Accessions" is annotation that is included with a submitted SNP (ss) when it is submitted to dbSNP. Submitter-Referenced Accessions are currently not included in the XML dump, but you can look up the accession using the ss number in the SubSNPacc.bcp table [ftp.ncbi.nih.gov/snp/organisms/chicken_9031/database/organism_data/SubSNPAcc.bcp.gz].(**1/5/06**)

Searching for SNPs Using Key Words or Names (Gene, Author, etc.)

I am looking for the corresponding rsID for a SNP that is commonly known as "human TNF alpha -308G>T" but can't find it in dbSNP. What can I do?

Unless the submitter used the exact name "human TNF alpha -308G>T "in their submission, it won't show up as a result in your dbSNP search. It is also possible that the authors who mentioned "human TNF alpha -308G>T" in their publication have never submitted this SNP to dbSNP.

Submission to dbSNP is voluntary, but we do try to encourage people to submit SNPs to dbSNP prior to publication so that they can cite their assigned dbSNP IDs (ss or rs) in their manuscript. Check the publication to see if the authors included a dbSNP ID for "human TNF alpha -308G>T" in the manuscript; if they have not, your other option is to BLAST the DNA sequence of human TNF alpha -308G>T (if you have it) against dbSNP to see if there are any matching rs numbers. **(11/17/05)**

I am trying to find the refSNP (rs) or submitted SNP (ss) numbers of SNPs described in the literature. The nomenclature used for these SNPs (IL1A[gene] -889C>T) is not supported by your search query.

Try the following search:

1. Go to the dbSNP home page. Once you are there, enter IL1A[gene] in the search box located at the top of the page.

2. On the resulting page, you will see 94 SNPs. Click on the "Limits" tab located near the top of the page.
3. Once you are on the limits page, scroll down until you find the “Observed Allele” section, and click on the box next to IUPAC code "Y" (Meaning C or T). Now, Click on the "Go" box located next to the search box at the top of the page.
4. The resulting page shows that by limiting your search this way, you have narrowed your search result down to 27 SNPs. **(3/7/05)**

When I search the dbSNP database for “microsatellite” with Limits of coding nonsynonymous and Homo sapiens, none of the records I retrieve contains the term “microsatellite”. Why?

Unless you specify the correct field tag, every field is searched, and every SNP record that has the word “microsatellite” anywhere in it will be returned. If you search using "microsat"[SNPCLASS] without limits set, you should retrieve 4954 records. You can use the Preview/Index option to get the correct field and value.

I tried a query using the keyword “mtDNA” and found 60 records, but many of them show SNP localization on chromosomes rather than in the mitochondrial genome. Why?

The SNPs that you retrieved using the mtDNA query are not true mitochondrial moltypes; they are incidentally related in some way with the keyword mtDNA. We only have mitochodrial moltype data for one organism, *Cooperia oncophora*.

If I search dbSNP using a gene name, would the search results include SNPs in the UTR?

The search may return SNPs in UTR regions if a locus has been defined to contain those regions on the genome.

The current Nature Genetics advance online papers indicates that David Page’s group submitted 95 new SNPs to dbSNP. How do I find these SNPs?

1. From the SNP home page, scroll down to the “Submission Information” section, and click on the words "By Submitter". This will take you to the Search/View Submitter Detail page.
2. In the “Search By” section of the Search/View Submitter Detail page, choose "Submitter name" and then choose "contains". Then type the name, "Page" (without quotation marks) in the text box, and click on the “Search” button.
3. Click on the name “Page” in the Handle column of the response page to go to the Contact Detail page.
4. Scroll down the Contact Detail page until you see a list of data batches submitted by the Page lab. To see a list of the SNPs you are interested in, click on the text “2006.01.23” in the "Submitter batch id" column to go to the “View SNP Submission Batch” page. The Submitted SNP (ss) ID numbers displayed on this page are links to the submitted SNP data. (**3/9/05**)

How do I interpret the results of a gene name query? I used the cyp2j2 gene to query dbSNP, but I am unable to determine the number of SNPs located within cyp2j2 by looking at the results page.

We have 40 human SNPs in our database associated with the cyp2j2 gene.
Start here:

1. Enter: cyp2j2[GENE] AND human[ORGN] in the Search box at the top of the form and click the Preview button.
2. The SNP count for the gene name queried will display as an active link that will take you to a list of available SNPs associated with the gene queried.
3. Try to explore the Help links on the left sidebar to learn more about Entrez SNP and the Preview/Index feature.

Why is it that when I search dbSNP for using "PRKAA1 Homo sapiens" I get 118 entries, but when I check the SeqView for them I find only 9 of the entries located within the sequence?

Seqview displayed the SNPs that are located on PRKAA1 mRNA. Most of the remaining SNPs are located in the intron and are not displayed.**(1/12/05)**

Retrieving Specific Data for a SNP

Finding CpG Island Data

Where can I get a graphic that contains sequence with CpG islands highlighted or bolded?

You can use Mapviewer to display and view CpG island locations, but at the present time, you will not be able to view CpG islands at the sequence level. To view SNP and CpG island locations:

1. Go to Mapviewer . Using the drop-down menus at the top, select on organism and gene symbol you wish to use in your search, and click on the “Go” button. For this example, I will use human and BRCA1.
2. Once the search for of your organism and gene symbol is complete, you will be taken to the “Genome View” page for your search. Scroll down the page until you see a list of “Map Elements” on the right. Select the Map Element of interest to you.
3. You will now be taken to a MapViewer page showing the Map Element you selected. Click on the “Maps & Options” button at the upper right-hand side of the page, to go to a window that will allow you to add maps and change display options.
4. To view SNPs and CpG islands, select “CpG Island” from the menu in the left side of the window, and click on “Add”. Then scroll down the list of maps, select “Variation” and click “Add”. At this point, you can add or remove other maps of interest, select which map will be the Master map, or change a number of other map options.

There is online documentation regarding CpG islands that you can review.
Other online CpG island viewers can be found by searching Google using the terms: "cpg island graphic viewer".**(3/22/05)**

Finding Flanking Sequence Data

Does dbSNP store flanking sequence for a given refSNP?

We don't store flanking sequence for a given refSNP (rs), but the flanking sequence we use for it is simple to get. We just use the sequence of the member-submitted SNP (ss) that has the longest flank. If the ss with the longest flank is in reverse orientation with the rs, we reverse the ss flank.

Finding Proximal (Neighboring) SNPs

How do I identify proximal SNPs located around a primary SNP?

You can find SNPs that are proximal to a given SNP by using the “Neighbor SNP” link located in the Integrated Maps section of the refSNP Cluster Report. For example, you can find the SNPs proximal to rs2515644 by scrolling down the cluster report until you came to the “Integrated Maps” section, which is located approximately half way down the page. Looking to the right, you’ll see the “Neighbor SNP” column. Click on the blue “View” links located in the column to view neighboring SNPs. (**10/13/06**)

Finding Functional (Synonymous/Non-synonymous, etc.) Information for a SNP

Is there a way of using dbSNP to determine which genes contain SNPs but do not contain non-synonymous SNPs?

You can search for genes without SNPs in the human genome, but here is no easy search for genes that have SNPs but no non-synonymous SNPs. One option is to download the SnpGeneReport from the dbSNP FTP site (the human genome_report sub-directory [ftp.ncbi.nih.gov/snp/organisms/human_9606/genome_report]) and filter the data. (**12/27/06**)

Is there any way I can query all of dbSNP to extract all known human SNPs for all C/T SNPs that cause non-synonymous changes in the coding region?

1. You can only query for all C/T SNPs that cause non-synonymous changes in the coding region by querying Entrez SNP:

1. Start on the Entrez SNP home page
2. Type "Human" (don't include the quotes) in the text box at the top of the page, and click on the “Go” box.
3. Now, click on the “Limits” tab located just a line below the Search text box to generate a form that contains many limit options.
4. Go to the , the “Functional Class” section, of the limits page and click on the box next to the words “coding nonsynonymous”.
5. Scroll down to the “Observed Alleles” section of the limits page and click the box beside “Y” (to set the limit for C/T). check box, and click on the “Go” box located next to the search text box at the top of the page.
6. The results for this search include 14404 SNPs. **(5/3/05)**

How do I search for non-synonymous SNPs in your database?

Conduct your search on Entrez SNP using the term “all[sb]” to retrieve all SNPs, then click on Limits and choose the filter(s) you want. Select the filter, coding nonsynonymous, under the Function class category.

Here's an example of such a search.

How do I download the functional (synonymous/non-synonymous/non-coding) information for a list of multiple RefSNPs?

You can use Entrez SNP to download the XML, ASN1, or Flat file (FLT) reports, each of which will show SNP functional class using a list of SNP ids (rs#). There are also programming utilities (eUtils) available to automate this process. The XML, ASN1, or Flat file reports for all SNPs in dbSNP are also available on the dbSNP FTP site [ftp.ncbi.nih.gov/snp]. **(2/14/05)**

How do I extract the following fields from dbSNP for a human non-synonymous SNP:

Get a flat file report that contains the information you want, with the exception of the frequency field.

Allele frequencies	Contig_position
Amino acid alleles	Orientation
Chromosome	Protein accession
Chromosome position	Position on protein
Contig accession	SNP alleles
Contig amino acid allele	SNP contig allele

Click on Limits and then select the following options: coding nonsynonymous and Homo sapiens. Then click Go. Select Flat File from the Display drop-down list.

How do I locate synonymous, non-synonymous, and frequency information for SNPs that occur within the coding region of each entry in the RefSeq database?

All SNPs are mapped to a RefSeq sequence. You can search for them using Entrez SNP and the following query:

1. Enter “all[sb]” into the Search box.
2. Select the following limits: Homo sapiens, coding nonsynon, coding synonymous, and validated by frequency.
3. Click Go.
4. Select the dbSNP Batch Report Display.
5. Select genotypeReport as the result and enter your email address. The result will be sent to you by email.

I do not understand the meaning of the "Function" category or any of its components (e.g. synonymous, contig reference, nonsynonymous)?

The NCBI handbook has documentation that should address your questions. Below is an excerpt from that document that should address your question:

“...[fxn-class] defines variation functional classes. We base class on the relationship between a variation and any local gene features. When a variation is near a transcript or in a transcript interval but not in the coding region, then we define the functional class by the position of the variation relative to the structure of the aligned transcript. In other words, a variation may be near a gene (locus region), in a UTR (mrna-utr), in an intron (intron), or in a splice site (splice site). If the variation is in a coding region, then the functional class of the variation depends on how each allele may affect the translated peptide sequence.

Typically, one allele of a variation will be the same as the contig (contig reference), and the other allele will be either a synonymous change or a nonsynonymous change. In some cases, one allele will be a synonymous change, and the other allele will be a nonsynonymous change. If any allele is a nonsynonymous change, then the variation is classified as a nonsynonymous variation. Otherwise, the variation is classified as a synonymous variation.

The Four Basic Outcomes When a Variation Is in Coding Sequence

- The allele is the same as the contig (contig reference) and hence causes no change to the translated sequence.
- The allele, when substituted for the reference sequence, yields a new codon that encodes the same amino acid. This is termed a synonymous substitution.
- The allele, when substituted for the reference sequence, yields a new codon that encodes a different amino acid. This is termed a nonsynonymous substitution.
- A problem with the annotated coding region feature prohibits conceptual translation. In this case, we note the variation class as coding, based solely on position.

Because functional classification is defined by positional and sequence parameters, two facts emerge: (a) if a gene has multiple transcripts because of alternative splicing, then a variation may have several different functional relationships to the gene; and (b) if multiple genes are densely packed in a contig region, then a variation at a single location in the genome may have multiple, potentially different, relationships to its local gene neighbors.” **(3/14/05)**

How do I determine the enzymatic activity of a catalase SNP in relation to the normal enzyme (i.e. is it 20% of normal, 40% of normal, etc.)?

dbSNP currently does not have functional data for the variant form of the protein. Try searching PubMed for published information.

This sort of information is difficult to computationally mine from the literature for millions of SNPs. Therefore, dbSNP will be instituting an online user annotation tool for the variation community to contribute functional data and references for SNPs in dbSNP. The annotation tool should be available sometime in the last quarter of 2006 (5/22/06).

Finding Linkage Disequilibrium Data

Can I search for linkage disequilibrium using the dbSNP page?

Currently, linkage disequilibrium cannot be searched from dbSNP. We may add this feature as we get more data from the HapMap project.

Finding Hardy Weinburg Probabilities

Where can I download population diversity data such as Hardy Weinburg Probabilities via FTP for all of the SNPs described in dbSNP?

The information you are looking for is located in the “genotype” file for the particular organism you are interested in. For example, if you are interested in population diversity data for human, go to the dbSNP FTP site, select the organisms [ftp.ncbi.nih.gov/snp/organisms] directory, then select the human [ftp.ncbi.nih.gov/snp/organisms/human_9606] directory, and then select the genotype [ftp.ncbi.nih.gov/snp/organisms/human_9606/genotype] file. The format of these files is described online. (**10/5/06**)

Where can I view the Hardy-Weinberg probability for rs35000?

To see the Hardy-Weinberg probability by population:

1. Scroll down the rs report page to the Variation Summary section and click on “Genotype Detail” located in the right-hand side of the section.
2. Once you are on the Genotype Detail Report page, scroll down to the bottom of the genotype list, where you’ll see the words “SNP Detail” located in a blue bar. Click on the “+” sign that is located next to the words “SNP Detail”.
3. To view the Hardy-Weinberg probability for the submitted SNPs clustered to this refSNP, click on the “+” sign located next to the refSNP of interest (in this case, there is only one: rs35000). **(11/2/05)**

Finding Hap-Tagged SNPs

How can I limit my dbSNP query to hap-tagged SNPs only?

dbSNP will not have hap-tag sets available until we start getting estimated haplotypes. Once we get estimated haplotypes, we must wait until we have "haplotype-tagged" sets before adding this filter in EntrezSNP.

We currently have 24211 refSNP(rs) that were used in defining haplotype blocks; most of these were submitted by Perlegen in Sep.2001, and the remaining 191 rs were from dbMHC's HLA project, which were also submitted in 2001. You can view the 24211 refSNP set in haplotype blocks by clicking on the words "Sample HapSet" and/or "Sample Individual" located in the Haplotype section of the blue left side bar on the rs page. **(4/12/05)**

How do I download a text file or excel file that contains SNPs rs numbers (5kb up and down stream), SNP locations (coding, intron, UTR, etc.), minor allele frequency (for Caucasian) and whether or not the SNP is HapMap tagged for our genes of interest?

Sorry, we don't have a data in the format that you requested. You can get the data as .xml (see instructions below) and parse out the fields of interest.

1. Search Entrez Gene for your genes of interest.
2. To the right of each gene displayed you will see the word “Link”. Click on this word and select “SNP” from the drop-down menu that appears.
3. Click the “Display” drop-down menu toggle located just underneath the tabs on the left near the top of the SNP page, to see a list of SNP display options. Select "XML" as the display type.
4. Click the drop-down menu “Send to" toggle located to the right of the “Display” toggle to see a list of destination options. Select “File” and save the result to your computer.

You can also look at the information using other reports you can select from the display drop-down menu (e.g. FASTA, Flat File, etc.) and aggregate the data from them. **(6/21/05)**

Finding Specific Hetrerozygosity Data

How do I get a list of SNPs and their heterozygosity reports in table format for the human SPP2 gene?

5. Search Entrez SNP using the [Gene Name] field.
6. Click on the tab marked “Human”, which is located in the second set of tabs at the top of the page.
7. Just above the “Human” tab you will find a menu of display options. Click on the blue arrow located in the “Display” text box to activate the drop-down menu. Select the “Chromosome Report” option.
8. Once the Chromosome Report option has been selected, you will be taken to the chromosome report format for your search results. If the average heterozygosity is available for your SNPs of interest, they will be displayed under the column entitled “avg het”, which is located toward the right side of the page. (**3/24/06**)

How do I find validated SNPs located in coding sequence that have heterozygosity over a certain threshold, say 30%?

5. Make a list of gene IDs (the gene ID for the example below is 4023) and look them up on Entrez Gene.
6. Upload the list of gene IDs on this page. Click retrieve to show the gene results.
7. Select SNP links and click the Display button to see the results.
8. Click on Limits at located at the top of the page, and select your filters (i.e., validated, coding, or heterozygosity).

Finding Organism Ploidy

Which dbSNP table/field contains organism ploidy?

dbSNP does not have a table for ploidy. To find the number of unique chromosomes in a particular species, go to the summary page and click on the organism to get to the NCBI taxonomy page. The number of chromosomes for a particular organism is found in the genome information section located in the middle of the page.

Finding Records with OMIM Data/Links

How do I get the title of an OMIM variation id?

We don't store OMIM titles in dbSNP. You'll have to look it up in OMIM’s morbidmap table [ftp.ncbi.nih.gov/repository/OMIM/morbidmap]. **(9/15/05)**

When I search for human SNPs and limit the results to chromosome 22, with non-synonymous coding and OMIM links, I got 346 records without OMIM links.

Your original query was (("Homo sapiens"[Organism] AND ("snp omim"[Filter] **OR** "snp structure"[Filter])) AND "coding nonsynonymous"[Function class]) AND 22[CHR]).

These request limits specify that the results must have OMIM, **or** must have structure; the results therefore, can have either OMIM or have structure — they don’t have to have both OMIM and structure.

If you change the query to: (("Homo sapiens"[Organism] AND ("snp omim"[Filter] **AND** "snp structure"[Filter])) AND "coding nonsynonymous"[Function class]) AND 22[CHR]), your results will include both OMIM and structure. (**3/29/06**)

For those SNPs that do have an OMIM ID number, what dbSNP report format will provide both a SNP and its specific OMIM ID number?

The OmimVarLocusIdSNP table [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/OmimVarLocusIdSNP.bcp.gz] contains the information you need for your organisim of interest (human, in this case). This table is located in your organism’s organism_data directory on the dbSNP FTP site.

Column definitions for this table are as follows:

ColumnDescription

1omim_id.

2The locus id the SNP is on
3 omim variation id.
4locus symbol
5Amino acid using the contig reference allele.
6Amino acid position in the protein.
7Amino acid of the snp variance.
8var class (used for internal dbSNP processing)
9snp_id (rs#)
Below is an extract from the OmimVarLocusIdSNP table, showing data arranged in columns from left to right in the order mentioned above:
1006502170001ALDH2E487K1671
102560710003ACTG1P332A111549200
102574890001ACTN3R577*11815739
1026801180001ADD1G460W14961
1027702700001AMPD1Q 12*117602729
1037201250002ADH2R369C12066702
(9/14/05)

For those SNPs that do have an OMIM ID number, what dbSNP report format will provide both a SNP and its specific OMIM ID number?

The OmimVarLocusIdSNP table [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/OmimVarLocusIdSNP.bcp.gz] contains the information you need for your organisim of interest (human, in this case). This table is located in your organism's organism_data directory on the dbSNP FTP site.
Column definitions for this table are as follows:
ColumnDescription
1omim_id.
2The locus id the SNP is on
3 omim variation id.
4locus symbol
5Amino acid using the contig reference allele.
6Amino acid position in the protein.
7Amino acid of the snp variance.
8var class (used for internal dbSNP processing)
9snp_id (rs#)
Below is an extract from the OmimVarLocusIdSNP table, showing data arranged in columns from left to right in the order mentioned above:
1006502170001ALDH2E487K1671
102560710003ACTG1P332A111549200
102574890001ACTN3R577*11815739
1026801180001ADD1G460W14961
1027702700001AMPD1Q 12*117602729
1037201250002ADH2R369C12066702
(9/14/05)

Finding Records for a Race or Specific Population

How do I search for known SNPs in several genes that occur only in the European/Caucasian population?

Although dbSNP does not have a classification for race and ethnic group, you can search on Entrez SNP for the gene and limit the subset to population class EUROPE. Enter the gene name or term in the search box, click Limits, and check the box for EUROPE under limit by Population Class.

How do I search for SNPs based on a set of criteria that may include race, allele frequency, etc.?

Try your search using Entrez SNP. It has search fields that are available, as well as some examples.
Filter your results using the limits found in Entrez SNP.

dbSNP is not allowed to classify SNP data based on racial or ethnic information, but you can filter or search SNP data using the field population class, which is based on geographic location.

Finding Sample Size Data

How do I find the number of samples (i.e. people) you've sequenced for a particular polymorphism of human maspin (serpin b5)?

dbSNP does not generate the SNP data. dbSNP is a depository for data submitted from hundreds of research groups. Each SNP may have a different sample size from another SNP. Search for rs2289519 using Entrez SNP. The results page for this search shows a row of colored buttons below rs2289519 that represent links. Click the pink "GeneView" button. This will take you to the "SNP linked to Gene" page, which shows that P176S corresponds to refSNP rs2289519. The total sample size for rs2289519 is located in the Population Diversity Section of the refSNP report.(**6/1/06**)

Finding Tab-Delimited Reports

How do I get a tab-delimited report of all mouse SNPs from dbSNP that would show refSNP id and fxn_class?

You can get the refSNP id and fxn_class from the tab-delimited bcp file mmSNPContigLocusId.bcp. mmSNPContigLocusId.bcp can be found in a file called mmNPContigLocusId.bcp.gz, which is located at the SNP FTP site. You can access the SNP FTP site the from dbSNP homepage sidebar. The definition of fxn_class is located in a file called SnpFunctionCode.bcp.gz, and the table columns are defined in dbSNP_main_table.sql.gz, which is located in the shared Schema [ftp.ncbi.nih.gov/snp/database/schema/shared_schema].

How can I download a tab-delimited file with Population Diversity information for hundreds of SNPs?

There isn't a tab delimited report format containing the data that you requested. You'll have to upload your list of SNPs to the batch query service to get the XML or ASN report and parse out the data you want.

Finding Validation Data

How do I get a flat file that contains the validation and strain information for all of the mouse strains in dbSNP?

There are two steps:

1. Go to Entrez SNP.
2. Click on Limits and choose Mus musculus in the organism section; choose validation status in the Validation section. If you are interested in several specific strains, you could enter the strain name in the search box to narrow the result.

Finding out Why a SNP has been Withdrawn

Why has rs11568324 been deleted from dbSNP? We, and other groups have genotyped this SNP, and although the minor allele frequency is very low (0.7%), we consistently found this SNP.

The rs11568324 cluster contained two submitted SNPs (ss#), both of which were from the same submitter. This submitter also submitted a "withdraw" request on September 7, 2006 for the SNPs in question.

The submitter has since re-analyzed their SNP data and has resubmitted their SNPs to us. The submitter indicated that their newly submitted SNPs may include some previously withdrawn SNPs. We are in the process of mapping the new SNPs from this submitter, and if any of these SNPs map to the position that rs11568324 used to map to, they will be assigned to that cluster.(**11/21/06**)

I searched db SNP several months ago for SNPs in SLC22A2 and its splicing variations, and found non- synonymous SNPs. I searched again today, and found that many of these SNPs have been deleted from dbSNP. Why?

The SNPs mentioned in your question were originally submitted by PHARMGKB (to see the submitter of a ss#, click on the ss#, which is located in the "NCBI Assay ID" column of the "Submitter Records" section of the refSNP report page). PHARMGKB has withdrawn 11740 of their submitted SNPs (ss#) for error corrections and will resubmit the data early next year (2007). PHARMGKB did not indicate the percentage of corrected SNPs that will have flanking sequence changes. (**12/19/06**)

Searching for SNP Genotype Data and Frequency Data

How do I code a URL so that a refSNP (rs) number can be linked directly to the genotype and allele frequency report?

There is a way to link the rs to the genotype and allele frequency report by using the Genotype server. Please look online for details. (**10/16/06**)

Genotype Data

After narrowing the range of my search for SNPs that are polymorphic between the C57BL6/J and 129 sv/j strains of mice by typing 3[CHR]105351425:107237468 [CHRPOS] into EntrezSNP, How do I get Entrez SNP to show me SNPs that are polymorphic between C57BL6/J and 129 sv/j?

Enter the following into Entrez SNP**:**

"Mouse[orgn] AND true[gtype] AND 3[CHR] AND 105351425:107237468 [CHRPOS]"

You could also use the genotype query form for this query, although using a genotype query will take some time since this is such a large region. **(11/17/06)**

How do I upload the data for only those SNPs that have genotype or frequency information?

There are two ways to upload the data for all SNPs that have genotype or frequency information.

The easiest way to do this is to upload the Genotype XML GenoExchange files. These files contain all dbSNP Genotype and allele frequency information arranged by chromosome for each organism. For example, the human build 125 genotype data [ftp.ncbi.nih.gov/snp/organisms/human_9606/genotype] are located in the genotype subdirectory of the human SNP directory. The schema for the GenoExchange files is also located online.

Another way to upload the data for all SNPs that have genotype or frequency information that requires a little more work, is to upload all the dbSNP database files, which are also available on our FTP site [ftp.ncbi.nih.gov/snp/database]. (**5/6/06**)

How do I find the genotypes for rs2074192 in dbSNP?

To view the genotypes of a single SNP:

1. Go to the SNP home page and find the "Search by ID" section. It is located below the announcements.
2. Type "rs2074192" (don't include the quotes) into the Search by ID text box, then click on the "Search" box. This will generate a refSNP cluster report for rs2074192.
3. Now, scroll down to the Variation Summary" section of the refSNP cluster report and click on the words "genotype detail" link located in the left hand side of the section. This will display the genotype report for this rs number. **(1/25/05)**

How do I download genotype and functional data in the XML format?

Genotype and functional data are stored as modules in two separate XML reports, so you'll have to download using eUtils and then combine the results from the two reports.

Below is an example of how to use eUtils to retrieve human SNPs that contain genotypes in the LPL gene:

1. eSearch using the term "LPL AND true[gtype] AND txid9606".
2. Parse the XML results to get a list of refSNP (rs) ID numbers.
3. eFetch to retrieve the reports for each SNP. (**2/7/06**)

I plan on using a Coriell panel of DNA samples from 50 individuals as controls in a genotyping study. Can I search dbSNP for single as well as multiple SNPs located in a single individual's DNA?

The dbSNP genotype report can be used to determine if a given SNP (or list of SNPs) has submitted genotypes for the samples you have purchased. Using the batch query option may be the best way for you to do this:

1. Go to the dbSNP home page and scroll down until you find the "Search" section heading on the blue left side bar, then click on the words "Batch Query", located four lines beneath it.
2. Scroll down to the bottom of the Batch Query page, and click on the "submission format" drop-down menu toggle to see a list of Batch Query submission options. Select "Enter RS#"
3. Now, enter your email address in the text box in the "Email" section of the page, select the appropriate organism from the drop-down menu in the "Organism" section, and enter your refSNP numbers along with their rs prefixes in the text area of the "Enter RS Numbers" section.
4. From the drop-down menu located in the "Select Result Format" section, select "genotype report". Click the "Submit" button.

An xml file containing all submitted genotypes for each of the refSNP numbers you entered will be emailed to you. The report will contain all populations and individuals typed for the SNPs you entered into the batch query since there is no option to select only specific individuals or populations at this time.

You can look at a schema document that contains details on the xml format for the genotype report.
You can also look at the SNPs one at a time by doing the following:

1. Go to the dbSNP home page and enter a rs# in the textbox at the top of the page, and click the "Go" button.
2. Click on the red "G" link shown in the resulting graphic to bring up the genotype report in html.
3. Click on the "+" sign next to a listed population name to expand a table of genotypes. In the tables, each row represents a sample. **(6/16/05)**

How can I get a dump of the new AFFY genotypes?

Go to the dbSNP homepage and click on By Submitter in the blue sidebar (under Search). Type "AFFY" in the textbox and then click the Search button. Affymetrix submissions will be displayed. Now, click on the submitter's handle to view the contact detail and obtain a list of the batches. Click on batch_id to see the populations. Finally, click on detail under view genotype to display the actual genotypes.

I found some sites marked as SNPs in the Mouse Ensembl viewer, but the mouse strain is labeled as "unknown". How do I find out which mouse strain has this polymorphism?

Go to the link Entrez SNP search link and click on the genotype report link (red letter G) on the graphic display. The strains are displayed on the genotype report under Genotype → Submitter's Id.

Frequency Data

How do I identify all dbSNP frequency data coming from the HapMap CEU population?

The HapMap Handle is "CSHL-HAPMAP" and the HapMap CEU population ID is "HapMap-CEU". Using the dbSNP "Population Detail" search, enter HapMap-CEU in the test box in the grey query section, and then select "submitter population ID" and "exact". Click on the "HapMap-CEU" link you get in your response to get the details for the HapMap-CEU population.

One way to get the allele frequency information for this population is to parse the genotype and allele frequency (genoExchange format) xml files found in the human genotype directory [ftp.ncbi.nih.gov/snp/organisms/human_9606/genotype] of the dbSNP FTP site. You can find documentation for the genoExchange format online.

All of the “ByPop” elements that have the attribute pop_id="1409" are from the HapMap CEU population.

If you are interested in finding allele frequency information for specific variations, or variations located in specific genes, SNPs, or particular regions, you may want to use the dbSNP’s genotype query. Users can specify genotype query output as XML, HTML or text. (**11/06/06**)

What is the expected frequency of eight SNPs occurring in the coding region of a single, 1,300 bp ORF?

The frequency of multiple SNPs is estimated to be two exonic SNPs per gene (coding and untranslated regions). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms is located online.

Suggestions for verifying your variations:

Sequence cDNA products from several independent PCR reactions or test multiple clones.

Use a different cell line for the control reaction.

How do I find SNP allele frequency information for different populations?

You can get SNP allele frequency for different populations by using the genotype report format. To get this format, select a refSNP from a population you are interested in, let’s use rs221 for this example.

1. Go to the Entrez SNP page and type in rs221 in the search text box at the top of the page, and click the “Go” box to generate the rs221 record.
2. Click the “Display” drop-down menu toggle located just underneath the tabs on the left near the top of the SNP page, to see a list of SNP display options. Select "genotype detail" as the display type
3. If you click on the “+” sign located just to the left of each genotype option, you will see SNP frequency for populations available on dbSNP. You can also get this data in xml format by choosing "genotype XML" as the display type in step 2. **(5/4/05)**

How do I access allele frequencies in dbSNP?

There are several ways to get allele frequency information:

- Use the average frequency information for a single reference SNP (refSNP, rs#).
- The frequency information for a single submitted SNP (ss#), is located toward the bottom of the page. The same data is also available in submission format.

To view genotype information for a particular population, do the following:

1. Go to the dbSNP homepage and select the Search by Population detail link and conduct a search as indicated in the documentation.
2. Once dbSNP generates a report of the population detail, select a sub-batch ID of interest to generate a Population Detail report.
3. To generate an Individual Genotype Batch report, select a batch ID of interest from the Population Detail report.
4. To view the individual genotypes of your batch, Select detail located under View genotype in the batch summary section. An example of the Individual Genotype Batch report is available.

Allele frequency data are available at dbSNP's FTP site in several formats—XML, asn.1, and submission format.

You can also obtain the average allele frequency for all rs numbers in the SNPAlleleFreq [ftp.ncbi.nih.gov/snp/database/organism_data/human_9606/SNPAlleleFreq.bcp.gz] table located in the organism_data directory for your organism.

How do I find population data on SNP frequencies posted on dbSNP? I am looking for large, random, ethnically defined population frequency data and do not have a PMID number.

Because you mentioned “PMID”, I’m assuming that you have a publication title or an author’s name. If so, you can search dbSNP using the publication title or the author’s name.

Clicking on the title leads you to the publication page. At the bottom, there is a list of all batches citing this publication. Here’s an example. Look for batch_type “frequency”.

The above approach will not give you all SNPs with large, random, ethnically defined population frequency data, but if you have a publication in mind, it will give you a start.

The “display” menu for dbSNP Entrez lists both "Allele Frequency XML" and "Genotype XML", even though these two selections give identical XML reports. Why have two reports showing identical information?

"Allele Frequency XML" provides only frequency data, while "Genotype XML" is more comprehensive, and reports both the frequency and genotype data if available. We provide the shorter version (Allele Frequency XML) for users that don't need the genotype data, so the reports will look identical if no genotype is available for the SNP. **(12/09/05)**

How do I find the populations used to determine SNP frequencies, and whether these frequencies vary by ethnic group?

Follow the example below:

1. Go to Entrez SNP
2. Type in "LPL" in the input text box at the top of the page, and click the "GO" button located to the right of the input text box.
3. You will get a report that has two sets of tabs at the top. Look at the lower set of tabs and click on "Human" tab.
4. You will see a list of rs ID numbers. Scroll down the list of rs ID numbers and look for one that has a red "G" located on the right end of the graph bar.
5. Click on the red “G”to show genotype details for that rs ID number. If the rs ID does not have a red ”G’, then no genotype detail is available for that rs ID

Another way of finding genotype information for an rs number is to use the "Genotype Detail" link located in the "Variation Section" of any refSNP report:

Follow the example below, using any refSNP page of interest.

1. Scroll down to the bottom of the report, where you will find the variation section.
2. Click on the blue “Genotype Detail” link at the bottom of the variation section
3. At the bottom of the list of genotypes, you will see "SNP Detail" bar. Find and click on the blue "+" link located at the front of this bar.
4. You will see frequency detail organized by each submitted SNP(within the refSNP cluster)and population.
5. If you select the "xml" format from the "Display" drop down list at the top of the page, you will also get the computed Hardy-Weinberg Probability **(10/04/05)**

Retrieving Large Numbers of SNPs at a Time (Batch Query)

Conducting BATCH Queries using Specific ID Types

Can I conduct a Batch query using Celera ID numbers?

The dbSNP batch query service now accepts hCV numbers. You can either enter the numbers in the text box entry site on the web, or you can upload a list from your computer.

1. Go to the dbSNP Batch Query site.
2. Scroll down to the text box with “Submission Format” in it, and click on the blue arrow to activate a dropdown menu.
3. Select “”Upload Celera ID” to upload a list from your computer, or select”Enter Celera ID” to enter the Celera ID numbers directly.
4. Follow the directions for entry given on the response page.

When you are ready to enter or upload your list, make sure that the list is formatted with one ID per line as shown below:

HCV1
HCV100
HCV100000
HCV1000000
HCV1000004
HCV100001
HCV1000012
HCV1000013
(**4/10/06**)

Retrieving Specific Data using BATCH Query

How do I get the chromosome, chromosome position, alleles and Ancestral allele for each SNP in a file of refSNP numbers I obtained from HapMap?

You can use dbSNP Batch Query service to download the Flat File Report format for a list of rs numbers. The Flat File Report contains chromosome, chromosome position, and allele information for each refSNP. You may also wish to look at the ancestral allele section of the SNP FAQ archive for further information on ancestral alleles. (**9/19/06**)

When conducting a BATCH query, how do I specify a specific set of fields for the query to retrieve (chromosome positions, rs#, and gene name)?

On the Batch query submission page select "Chromosome Report" as the result format. (**4/10/06**)

How do I retrieve all the reported SNPs, their frequencies, function, chromosome positions and FASTA files for 50 genes? Would I use a Batch search? If so,how?

The Batch query is not the correct way to proceed. Instead, you will have to write a program using the eutils programming utilities:

1. Use eSearch to perform the search, and then parse out the refSNP (rs) numbers you need. Here are the results for a search for SNPs in the human LPL gene.
2. Use eFetch to retrieve the SNP reports (**6/20/06**)

Is there a way I can download the functional information for a list of about 60,000 SNPs in text file format?

You can upload your list of refSNP numbers (30,000/per load) to the batch query service and request the “FLATFILE” report.

Click this link to see the FLATFILE report for rs1855025. **(5/12/05)**

How do I get SNP frequency information using a batch query?

Frequency information is located in the genotype report. Please use the batch query and select genotypeReport.

How do I use dbSNP refSNP (rs) numbers to get a list of chromosome positions?

1. Go to the dbSNP home page, and scroll down to the “Batch” section.

2. In the “upload list” sub-section, select “Reference SNP ID (rs)”.

3. On the data request form, enter the email address where you want your data sent, select the organism of interest from the “organism” dropdown menu, and in the “Select Result Format” section choose “CHROMOSOME RPT” from the dropdown menu (3/1/06).

How do I use dbSNP batch query to convert a series of SNPs from a local lab to rs and ss numbers?

Create a list using only the local SNP ID, but don't include the handle.

Do you have software that will tell me if a large number of SNPs (obtained from a non–dbSNP source) are either typed or part of the HapMap project list?

You can use Entrez SNP to filter out those SNPs that are part of HapMap project list, but as you have a large number of records as input, go to Batch Entrez and follow the second set of online instructions for accession list input.

The list of rs numbers you upload to Batch Entrez should contain just the numbers, not the rs prefix, as in the example below:

13054637
13009028
12843774
12783966
12779960

1. Upload your list of rs numbers according to the instructions, and click the “Retrieve” button, located at the top of the page near the “File” text box.

2. When the result of your search is displayed, click the "Preview/Index" tab, located in the first set of tabs just below the text box at the top of the page.

3. Go to the “Most Recent Queries” section of the Preview/Index page, and click on the search number that corresponds to the list of ID numbers you uploaded. When you do this, a pop-up menu will be displayed. Select “AND” from the list of options in this menu.

4. Apply the HapMap filter by using the following steps:

   - Go to the “Add Terms to the Query or View Index” section of the Preview/Index page and find the text box with the words “All Fields” in it. Click on the blue arrow to the right of these words to activate a drop-down menu. Select the word “Filter” from this menu, and click on the Index button, located to the right of the drop-down menu.

   - Select "hapmap" from the generated filter terms.

   - Click on the "And" button, located just above the list of filter terms to add the hapmap filter to your query.

   - Click on the "Go" button, located at the top of the page next to the text box.

   - The resulting page will show you the HapMap SNPs contained in your submitted list.

5. To down load the filtered list to your computer, go to the Display drop-down menu located near the top of the page between the two sets of tabs, and click on the blue arrow to the right of the words “Graphic Summary” to activate the menu. Select the “UI List” option.

6. Located to the right of the “Display” text box, is the “Send To” drop-down menu. Click on the blue arrow to the right of the words “Send to” to activate the menu. Select “File” from the list of options, and when prompted, enter the file the data is to be sent to. (**4/19/06**)

Can you send me the FASTA sequence for a list of rs numbers?

You can use dbSNP Batch Query Service to download the FASTA format for a list of rs numbers. (**9/19/06**)

Using BATCH to Retrieve Unclustered Submitted SNPs (ss)

How can I get information for a large number of submitted SNPs (ss) whose refSNP (rs) numbers have yet to be assigned, and will therefore not generate XML or FLATFILE reports in response to a Batch query?

Use the following instructions to upload the ss numbers of interest using the Batch query service and request the Submitted SNP “SubSNP” details report format:

1. Go to dbSNP Batch Query page.
2. Scroll down to the submission format drop-down menu and select “Upload SS#”.
3. Enter your email address in the appropriate text box, and select the organism of interest from the organism drop-down menu.
4. Select "SubSNP Details" from the Result Format drop-down menu.
5. Upload the series of ss numbers of interest using the “Browse” button. Once you have uploaded the ss numbers, Click on the “Submit” button.

   The Batch query limit is 30,000 SS numbers per request. (**4/4/06**)

BATCH Reports compatible with Excel

Is it possible to convert a BATCH report into an Excel spreadsheet? I want to use the Excel search function to query using a refSNP (rs) number so I can easily find the gene name, variation, function, and frequency information for the refSNP.

When you are entering your BATCH query, select the chromosome report as your result format. The chromosome report format is a tab-delimited table that can be imported into Excel. You’ll have to use efetch to get the allele frequency from the XML files (FREQXML), however. Remember that NOT all SNPs have frequency data. Examples of the different dbSNP formats retrieved using efetch located online. (**7/28/06**)

BATCH Queries and XML Formatted Reports

How do I extract data from XML output for rs batch queries?

I would recommend that you use an XML parser because of the complex structure of XML. There are parsers available for most common computer languages. You can see an example of parsers used in PERL. Another option is to use XSLT to extract and transform XML to the format of your choice.

When using XML output for rs batch queries, how do I select the frequency data for specific rs numbers from among the frequency data listed for all the ss numbers?

To get the rs allele frequency directly, use the FTP file, SNPAlleleFreq.bcp.gz, located in your organism’s organism_data directory (ex: human [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data]).

The fields in the file are tab delimited and are defined in the schema table for your organism (ex: human [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/schema/human_9606_table.sql.gz]).

Why do we find rs386014 displayed in both HLA-A’s and HCG4P6’s batch report?

rs386014 is annotated to HLA-A on the Celera assembly using the gene model (contig---> mRNA---> gene). The same SNP is associated with HCG4P6 because the variation has been mapped within 2 kb of an mRNA transcript for the HLA complex group 4 pseudogene 6.

(**1/3/06**)

If I need to get more than 40 rs numbers from XML-formatted genotype and allele frequency reports, will I need to write a Perl script to do so?

Here are instructions for obtaining an genotype exchange XML file for a set of specific SNPs:

If you have a list of SNPs (rs numbers), you can upload the list to batch query by using the rs page. At the rs page, type in your email address. Then select genotypeReport as the format and click on submit. The report will be emailed to you.

Alternatively, if you want a large list of rs numbers and want to start from a Entrez query, do the following:

Once you have narrowed your Entrez query to the list that you want, select the dbSNP Batch report option, click display, and follow the instructions in the rs list example.

In addition to the SNP-specific queries, we are currently exploring various options for allowing users to specify population, pedigree, and/or individual filters as limits to the downloads for the genotype exchange XML report.

Sequence Data provided in BATCH Reports

Will the results for a BATCH query that I've submitted contain SNP flanking sequences from various builds or will the flanking sequences be from the current build (b125)?

The flanking sequences will be from the current build — dbSNP build 125, which is based on human genome build 36.(**5/10/06**)

BATCH query Errors

I have a batch query I want to access but cannot gain entry to the FTP server to retrieve it.

The file you're trying to access was created in 2003 and has been deleted from the system. We cannot keep files in the FTP site for than 48 hours because of limited resources. Please submit your query again to generate a new file.

I've tried to batch download pig SNPs recently added to dbSNP but couldn't get anything besides a list of the SNPs to appear on the dbSNP webpage, and the "Batch Query Result" email attachment I was sent was empty. Why?

The batch query service only dumps reports for SNPs that have been clustered and assigned with reference SNP (rs) numbers. The pig SNPs you are inquiring about have yet to be clustered and assigned with reference SNP (rs) numbers.

When I enter a series of submitted SNP (ss) numbers into a Batch query, select *Homo Sapiens* selected as the organism, and select FLATFILE or XML as the report format, the query generates no information.

Only submitted SNPs that have had refSNP (rs) numbers assigned to them can be as can be released in the FLATFILE or XML report formats. The submitted SNP (ss) numbers you submitted for a Batch query were only just recently assigned to some new submissions that came into dbSNP, and therefore have not been assigned refSNP (rs) numbers yet. The rs numbers will be assigned by the next dbSNP release. (**3/29/06**)

I am unable to download Batch ID: AFFY_071103 because the file is too large. How do I download it?

To download the genotypes for the Batch ID: AFFY_071103, please follow these instructions:

1. Go to the dbSNP homepage and select the New Batches link located under the Submission Information section.
2. Fill out the form as follows: organism="Homo Sapiens-9606", Batch type="Variations with individual genotype data (IND)", Submitter Handle ="starts with" ="AFFY". Then click on Search Batch.
3. Select the link "AFFY_071103" under Local Batch ID.
4. Click on the download, all sort by button and click go. This will allow you to download the genotype information to your PC in a zipped file that is ~7 MB. Then use Winzip to unzip it. Unzipped, this file will be ~231 MB.

Our email server strips away emails that contain attachments ending in .zip or.gz. Is there another way to retrieve batch reports?

You can use dbSNP's Entrez batch query system if you're querying with refSNP (rs) numbers. The Entrez SNP batch query system is fast, but it can accept only rs numbers, and as such you need only submit the number without the "rs" prefix attached.

1. Go to the SNP batch Entrez page
2. Click on the "Browse" button at the top of the page, and select the rs list you intend to query.
3. Click on the "Retrieve" button located next to the “Browse” button
4. When the result is shown, select the “Display” option and select a report from the list of available report types.
5. Select the “Send to” option to save or print the report.(**5/19/06**)

I downloaded 30 submitted SNP (ss) accessions from the SubSNPAcc.bcp.gz file and submitted them as a batch query, but I have yet to receive a fasta file back.

The batch query does not work for unclustered SNPs, and since the records you submitted represent submitted SNPs (ss) that were not mapped and/or clustered, you did not receive a response. You can download the records for these SNPs from the dbSNP FTP site [ftp.ncbi.nih.gov/snp/organisms/macaque_9544/ss_fasta/2005]. (7/18/06)

Searching dbSNP using Sequence Data Homology (Search using BLAST)

Can I search dbSNP for sequences using BLAST?

The dbSNP BLAST search is located on the left side bar of the dbSNP homepage, or you can get there directly.

Is it possible to search dbSNP using a sequence rather than an accession number or rs number?

Yes. You can align your sequences against those in the human SNP database by using a program called BLAST. Click on BLAST SNP on the SNP homepage side bar, which brings you to SNP BLAST by chromosome. SNP variations are encoded using IUPAC notation in the BLAST database.

Here is an example of the BLAST output**:**

```
Query: 241 gtgcttcctgggctccctggctgtgctgctgtgtgtgtgcacggagcgtgtgcagtacta
           ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 313 gtgcttcctgggctcsctggctgtgctgctgtgtgtgtgcacggagcgtgtgcagtacta
```

You can see in the BLAST output above that there is an existing variation in dbSNP (note: the IUPAC notation “s”, shown above, represents g or c).

How do I determine if SNPs I’ve discovered have already been reported to dbSNP by another lab?

Align your sequences against those in the human SNP database using a program called BLAST.

Click on BLAST SNP on the SNP homepage side bar, which brings you to SNP BLAST by chromosome.

SNP variations are encoded using IUPAC notation in the BLAST database. If the position of your variation does not match any of the SNP positions as indicated by the IUPAC notation, then you have a new SNP.

Example of BLAST output:

```
Query: 241 gtgcttcctgggctccctggctgtgctgctgtgtgtgtgcacggagcgtgtgcagtacta
           ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 313 gtgcttcctgggctcsctggctgtgctgctgtgtgtgtgcacggagcgtgtgcagtacta
```

You can see in the example BLAST output above that there is an existing variation in dbSNP (note: the IUPAC notation “s”, shown above, represents g or c).

Can you give me some advice on sending queries from BLAST (URL) API to your database?

Did you mean URL BLAST API or C/C++ BLAST api? You could use BLAST WEB API to BLAST directly against dbSNP. Have you looked at the User’s Guide?

I've tried BLASTing dbSNP in *Homo sapiens* chromosome 6 using a sequence that contains a known SNP but get no sequence comparison results—just the RID number (1094764833-2783-170044276144.BLASTQ4).

I noticed that you included your BLAST Request ID (RID) in your question. GOOD!

It is always a good idea to save your Request ID (RID) when it first appears at the top of the format page, which is displayed after you select Submit. The RID will allow you or anyone else to retrieve the formatted BLAST result of your query without redoing the time-consuming BLAST.

I used your RID number to retrieve your BLAST output and found that the alignment was indeed returned. The fact that you did not see an alignment is puzzling. It could be that you possibly did not have the correct formatting options checked on the format page, but I think they are set to sensible defaults automatically.

Please try this to view the results of your 196-base example using the RID you provided and the following instructions:

1. Go to the main BLAST page and click on Retrieve results by RID under the word Meta, which takes you to the BLAST Request ID page.

2. Now, enter your request ID, 1094764833-2783-170044276144.BLASTQ4, in the format options box. Please make sure that the number of descriptions and the number of alignments are set greater than zero. Now, click on format.

3. You should see the page displayed. If this doesn't work, please let us know. If it does, try repeating your query to dbSNP, copy your RID somewhere just in case, and click on "format". If it doesn't format, use your browser to go back to the previous format page and repeat these steps as needed until output appears.

4. Make sure that the formatting options are sensible, and that the number of descriptions and the number of alignments are set greater than zero. Also, indicate that the alignment should be shown in HTML ("show ! alignment! in !HTML!"). The alignment view should probably be set to !pairwise! unless you know better.

We are in the process of developing SNP web pages, and was wondering if you have a URL for SNP BLAST that includes the parameters (sequence, organism) already filled in.

The best method is to use URL API, and the best documentation for using this is on the NCBI Learning Center website. Once you are on the Leaning Center website, scroll down the list and select "URLAPI (updated)", near the bottom of the page.

Toward the top middle of the page there is also a link called "Remote accessible BLAST databases". Please note, however, that some of the contents of this document are out of date.

Example:

To call the zebrafish snp database, use the following text in your URL:

&DATABASE=snp/zebrafish_7955/zebrafish_7955&

Other BLAST parameters are specified on the 'URLAPI (updated)' page.

If a large query sequence, or a multi-FASTA file, is to be used as the query, see the"Sample Perl Script" page. (**4/19/06**)

I want to download the current set of sequences in dbSNP to format for a local BLAST search. Could I use a "wget --mirror" to pull down the contents of /snp/organisms/*/ss_fasta/*/*", and then concatenate the files?

We already have the files you need on the dbSNP FTP site as FASTA files. For example, to get human ss FASTA data [ftp.ncbi.nih.gov/snp/organisms/human_9606/ss_fasta], go to the human organism directory, and select ss_fasta,and then select the year in which the data you need was submitted.

You can also get rs FASTA organized by chromosome [ftp.ncbi.nih.gov/snp/organisms/human_9606/rs_fasta] in the dbSNP FTP site, or you could also blast dbSNP rs sequences directly online.

For more information, please check the dbSNP handbook.

(**9/13/06**)

Searching for SNPs not yet Submitted to dbSNP

How do I search dbSNP to determine if some SNPs and deletion/insertion variations I have identified are novel?

Try BLASTing dbSNP with your sequence. To do this, go to the SNP home page, locate the “Search” section in the left-hand side bar and click on "Blast SNP".**(3/8/05)**

How do I determine if a sequence I have contains any of the variations housed in dbSNP?

You can try to BLAST dbSNP by clicking on the BLAST SNP button located on the dbSNP homepage sidebar, which will take you to SNP BLAST by chromosome.

How do I determine if SNPs I’ve discovered have already been reported to dbSNP by another lab?

Align your sequences against those in the human SNP database using a program called BLAST.

Click on BLAST SNP on the SNP homepage side bar, which brings you to SNP BLAST by chromosome.

SNP variations are encoded using IUPAC notation in the BLAST database. If the position of your variation does not match any of the SNP positions as indicated by the IUPAC notation, then you have a new SNP.

Example of BLAST output:

```
Query: 241 gtgcttcctgggctccctggctgtgctgctgtgtgtgtgcacggagcgtgtgcagtacta
           |||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 313 gtgcttcctgggctcsctggctgtgctgctgtgtgtgtgcacggagcgtgtgcagtacta
```

You can see in the example BLAST output above that there is an existing variation in dbSNP (note: the IUPAC notation “s”, shown above, represents g or c).

How are the contig and chromosome positions of SNPs determined? I found some new polymorphisms and would like to determine these positions.

Here at dbSNP, we map refSNP sets ourselves using MEGABLAST. Most commonly, the contig position of a polymorphism is found when the SNP is mined computationally. For instance, the SSAHASNP program mines SNPs from overlapping traces of the clones used in the genome assembly.

Please look at this example.

Three of the submitter records for refSNP cluster rs3736544 were computationally mined (see those with SSAHASNP in the Handle/Submitter ID field), and the contig coordinates were used to construct the submitter ID.

If you have already sequenced the flanks surrounding your polymorphisms, you can use the form on the contig BLAST page to determine the contig coordinates.

If you have a large number of contigs, it is more practical to download the entire genome sequence genome [ftp.ncbi.nih.gov/genomes/H_sapiens] and BLAST them yourself.

Searching for SNPs in Raw Sequence Data

Can NCBI provide me with a tool that will aid me in identifying SNPs from raw sequence data?

NCBI doesn't have a tool for identifying SNPs, but the Sanger institute offers a polymorphism detection tool called “ssahaSNP” for academic use, and you can always try searching PubMed for publication information on SNP detection software. (**1/30/06**)

Searching for SNPs Using EST Alignments

Do you know if there are tools for SNP discovery in the absence of genome sequence using multiple alignments of ESTs?

Gabor Marth wrote the program PolyBayes for that specific task. dbSNP has implemented PolyBayes as an internal tool for mining polymorphisms in genome pipeline activities. If you would like to use PolyBayes for a research project or within a research group, you should request both the program and a license to use the program from Gabor Marth. His contact email is marth@bc.edu**(2/10/05)**.

Searching for SNPs Between Mouse Strains

After narrowing the range of my search for SNPs that are polymorphic between the C57BL6/J and 129 sv/j strains of mice by typing 3[CHR]105351425:107237468 [CHRPOS] into EntrezSNP, How do I get Entrez SNP to show me SNPs that are polymorphic between C57BL6/J and 129 sv/j?

Enter the following into Entrez SNP**:**

"Mouse[orgn] AND true[gtype] AND 3[CHR] AND 105351425:107237468 [CHRPOS]"

You could also use the genotype query form for this query, although using a genotype query will take some time since this is such a large region. **(11/17/06)**

I want to find SNPs in the 129 and Bl6 mouse strains that are the same as each other but differ from BALB/c using the Search Mouse SNP between strains tool. I stipulated that BALB/c was to differ from 129/Sv, 129X1/SVJ, 129X1/SV, C57BL/6, C57BL/6J, and *Musmusculus* C57BL/6J. This query returns extra SNPs that do not conform to my requirements.

The Search Mouse SNP between strains tool will return any SNPs that are different between strain BALB/c and strains 129 or C57BL. It does not work for your query where SNPs have to be common between strains129 and C57BL and different from BALB/c.

The way to get around this problem is to perform two searches:

First, perform a search using the Search Mouse SNP between strains tool, where you ask for those SNPs that are different between reference strain BALB/c and strains 129/Sv, 129X1/SVJ, 129X1/SV, C57BL/6, C57BL/6J, and *Mus musculus* C57BL/6J. Then perform a second search using the Search Mouse SNP between strains tool, where you ask for those SNPs that are found in common between reference strain 129/Sv and strains C57BL/6, C57BL/6J, and *Mus musculus* C57BL/6J.

Save the results from each search file in the brief format, and then filter for the SNPs (rs#) that are in both files.

Searching for SNPs contained in the NCBI Genome Build

How can I get the SNP data for NCBI's genome build 34?

To get the mapping coordinates for build 34.3, please see the SNPContigLoc_34_3.bcp.gz [ftp.ncbi.nih.gov/snp/database/organism_data/human_9606/b125_SNPContigLoc_34_3.bcp.gz] table for the species of interest on the dbSNP FTP site.

You might also wish to consult tables b125_SNPMapInfo_34_5 and b125_ContigInfo_34_5. in the same FTP directory. Please see the Data Dictonary for table descriptions of the above tables.

Please note that some refSNP (rs) numbers might have been merged (if they are found to map the same location at a later build). The rs merge history is located in the RsMergeArch table, located in the same FTP directory as the aforementioned tables. (**4/28/06**)

Searching for Data Histories and Historical Data in dbSNP

Could you tell me the number of validated and unvalidated SNPs in dbSNP build 120?

Build120 was released on Mar 16 2004 12:00AM and mapped to the NCBI Human Genome 34 ver. 3. Build 120 has a total of 9,098,790 human SNPs, and 4,267,639 of those are validated.

Please note that dbSNP validation status is delineated by various levels. You can see definitions for the various validation status levels by clicking on the words "Validation Status" located on any RefSNP page in the table header just below the banner that reads “Submitter records for this RefSNP Cluster”. **(3/14/05)**

Where can I find the history of dbSNP mouse submissions — specifically those from Celera, as well as other commercial and non-commercial labs?

1. Go to the dbSNP home page.
2. Locate the Submission Information section and select the “New Batches” link.
3. Select “mouse_10090” from the organism dropdown menu and then choose Select “Variation submissions (SNP)” from the Batch Type dropdown menu
4. Click the Search Batch button at the bottom of the page, and you will be given a list of SNP batch submissions that includes submitter handle, batch ID, and the number of variations submitted in descending calendar order. (**2/21/06**)

How do I query dbSNP so that it will return a flat or xml file containing the new RefSNP (rs) ID number into which a previously valid rs recently merged?

You can get the rs merge history of all rs numbers from your organism’s (human in this case) RsMergeArch table [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/RsMergeArch.bcp.gz] located in on the dbSNP ftp site.

The following example shows that rs4344934 has been merged to rs1107123:

gzcat RsMergeArch.bcp.gz | grep 43449344344934 1107123 123 1 2004-09-24 18:49:00 2004-10-10 11:55:00 1107123 1

(5/25/05)

Some RefSNP (rs) numbers in dbSNP are merged into one rs number. Where does dbSNP provide the merge history?

The ftp directory has been re-arranged after dbSNP was split into a set of organism specific databases. The RsMergeArch table [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/RsMergeArch.bcp.gz] for human SNPs is still located in the dbSNP FTP site [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data], it’s just in a new file. **(11/8/05)**

How do I locate information for two fairly old submitted SNPs: ss35527513 and ss35527511?

Search for these ss numbers by using the “Search by ID” query module located on the dbSNP homepage:

1. Enter the ss ID number in the input text box.
2. Select "NCBI Assay IDs" from the drop down menu to the right of the input text box.
3. Click “Search”
4. You will get a “Submitted SNP(ss) Report” for the ss ID you entered for your search.
5. Click on the blue ss Id number located at the top of the report to go to the submitted SNP page for this ss ID number, or click on the blue rs Id number also located at the top of the report to get refSNP mapping information. **(10/14/05)**

Finding a Build Number Based on a Date and Vice Versa

I downloaded dbSNP on Jun 6, 2004, and now need to provide a reference for this build in a manuscript. Can you tell me the build number?

It looks like it was dbSNP build 121. If you scroll to the bottom of the SNP summary page, you can query there for release dates from the most recent build all the way back to build 106. **(1/19/05)**

Where do I find the dates when a particular SNP was released for public view?

You can find these dates by checking the build history page for the release dates (these immediately follow the submission dates. PLEASE NOTE that the dates found in the NCBI databases are not to be used for patent purposes. See the NCBI Copyright and Disclaimer page for additional information.

Searching for dbSNP Statistics

What percentage of SNPs deposited in dbSNP is located in coding regions (exons) versus non-coding regions (introns + intergenic regions)?

You can get these numbers by querying Entrez SNP:

1. Type "All[sb]" (without the quotes) in the search text box located at the top of the page.
2. Click on the “Preview/Index” tab located just below the search text box.
3. Scroll to the bottom of the Preview/Index page, and click the drop-down menu toggle arrow just to the right of “All Fields”. Select "Function Class" from the list.
4. Click on the "Index" box.
5. A list of SNP functional classes will appear with the number of SNPs in each functional class located to the right of each functional class. **(6/21/05)**

How do I find the number of non-synonymous SNPs in dbSNP?

To get the number of non-synonymous SNPs, click on the geneReport link located next to HOMO SAPIENS under Build Statistics on the build summary page.

How do I find the number of rsSNPs that map uniquely to the human genome?

Conduct the query on Entrez SNP:

1. Enter the following query in the Search box to search for human SNPs that map once to the genome and not on unknown chromosomes:
2. (human AND 1[WEIGHT]) NOT un[CHR]
3. Change the query weight to 2 for the SNPs that map twice.

How do I determine the number of SNPs there is per gene across all genes in the human genome?

Use the ContigLocusId table, which contains all the SNPs mapped to genes in the human genome. The ContigLocusID table is located in the organism data directory [ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data] of the human database in the bSNP FTP site. (9/2/05)

Errors that Occur While Searching

I'm using GenBank to analyze the variations in *Rattus norvegicus*. When I double-click on a specific variation icon, I get the following message: “rs8149781 is not in dbSNP”. The rs number is correct, so how is this possible?

The annotation on the *Rattus norvegicus* assembly is temporarily more current than the dbSNP public webpages. As soon as NCBI human assembly build 34 is made public, we will release dbSNP build 117. Once that happens, the dbSNP links in the GenBank annotation will be active.

Sorry for the confusion. dbSNP is working toward a more flexible data structure that will allow us to release public data in a more timely fashion so that these synchronization issues can be resolved.

Why does the “Common Query Filters” link located on the left blue side bar of the Entrez SNP page lead to the "My NCBI page"?

The link is correct. My NCBI provides a link to commonly used SNP filters, or allows you to set up custom filters for yourself. You will need to set up an account and sign in to "My NCBI" to access the filter resource. **(12/9/05)**

Is there a bad link to the SNP records for NRP2 (human neuropilin 2)? When I key RP2 into the Entrez SNP search box, it takes me to NELL2, a completely different gene.

By keying only “RP2” into the search box, you are telling Entrez to search not just the gene name field but all fields that contain “RP2”; so you end up with 1465 records, most of which do not pertain to your gene of interest. If you key in “RP2[Gene Name]” into the Entrez SNP search box, then Entrez SNP will retrieve only those SNP records that contain RP2 in the “Gene Name” field (352 records).

How do I interpret the results of a gene name query? I used the cyp2j2 gene to query dbSNP, but I am unable to determine the number of SNPs located within cyp2j2 by looking at the results page.

We have 40 human SNPs in our database associated with the cyp2j2 gene.

Start here:

4. Enter: cyp2j2[GENE] AND human[ORGN] in the Search box at the top of the form and click the Preview button.
5. The SNP count for the gene name queried will display as an active link that will take you to a list of available SNPs associated with the gene queried.
6. Try to explore the Help links on the left sidebar to learn more about Entrez SNP and the Preview/Index feature.

I am using Entrez SNP on Internet Explorer and get an error message that reads: “A runtime error has occurred. Do you wish to debug? Line:71 Error: 'undefined' is null or not an object [Yes] [No].” This occurs during searches using any chromosome number.

You must have tried to get access while the binary applications were being updated. You should change your URL to specify the [CHR] field to search for SNPs by chromosome. Here's an example.

Additional fields and examples are shown on the Entrez SNP site.

When I search for human SNPs and limit the results to chromosome 22, with non-synonymous coding and OMIM links, I got 346 records without OMIM links.

Your original query was (("Homo sapiens"[Organism] AND ("snp omim"[Filter] **OR** "snp structure"[Filter])) AND "coding nonsynonymous"[Function class]) AND 22[CHR]).

These request limits specify that the results must have OMIM, **or** must have structure; the results therefore, can have either OMIM or have structure — they don’t have to have both OMIM and structure.

If you change the query to: (("Homo sapiens"[Organism] AND ("snp omim"[Filter] **AND** "snp structure"[Filter])) AND "coding nonsynonymous"[Function class]) AND 22[CHR]), your results will include both OMIM and structure. (**3/29/06**)

Why is it that when I search dbSNP for using "PRKAA1 Homo sapiens" I get 118 entries, but when I check the SeqView for them I find only 9 of the entries located within the sequence?

Seqview displayed the SNPs that are located on PRKAA1 mRNA. Most of the remaining SNPs are located in the intron and are not displayed.**(1/12/05)**

If I query all fields of dbSNP for a submitted SNP (ss) number that I submitted using the home page default search, I do not get a result, yet if I query a an older version of the dbSNP home page, the ss number I queried for is found.

Although your SNP was submitted in the latest build cycle, it missed the refSNP (rs) clustering process of the new build (in this case b125), which means your SNP has an ss number, but no rs number. You couldn't find your ss number because Entrez SNP indexes only refSNP (rs) numbers. Also, B125 had numerous internal BLAST pipeline changes; so many ss numbers were not clustered. In the future, we hope to get all ss numbers clustered before a new build — so this should not be a problem again.(**1/4/06**)

BATCH Query Errors

I have a batch query I want to access but cannot gain entry to the FTP server to retrieve it.

The file you're trying to access was created in 2003 and has been deleted from the system. We cannot keep files in the FTP site for than 48 hours because of limited resources. Please submit your query again to generate a new file.

We have mouse SNPs with Celera IDs, but when we BLAST these sequences, we cannot find any corresponding rs numbers in the gene in question. How do we find these SNPs and how do we report them?

The mapping of SNPs to a gene is completed by BLAST analysis; the mapping is therefore computed rather than experimental. As a result, some submitted SNPs (ss) that have low complexity sequence may therefore map to locations on the genome other than the location at which you are looking.

SNPs are also mapped to multiple assemblies (i.e. reference and Celera), and since these assemblies are different, a SNP may map to different genes within the different assemblies. For example, rs13186575 is mapped to "HSPD1" on the Celera assembly and to "CDH12 on the reference assembly.

Some SNPs have discordance between their computed positions and their experimentally observed positions. We are planning to develop online tools that will allow users to provide annotations and corrections when such discordance arises. (**3/3/06**)

I've tried to batch download pig SNPs recently added to dbSNP but couldn't get anything besides a list of the SNPs to appear on the dbSNP webpage, and the "Batch Query Result" email attachment I was sent was empty. Why?

The batch query service only dumps reports for SNPs that have been clustered and assigned with reference SNP (rs) numbers. The pig SNPs you are inquiring about have yet to be clustered and assigned with reference SNP (rs) numbers.

When I enter a series of submitted SNP (ss) numbers into a Batch query, select *Homo Sapiens* selected as the organism, and select FLATFILE or XML as the report format, the query generates no information.

Only submitted SNPs that have had refSNP (rs) numbers assigned to them can be as can be released in the FLATFILE or XML report formats. The submitted SNP (ss) numbers you submitted for a Batch query were only just recently assigned to some new submissions that came into dbSNP, and therefore have not been assigned refSNP (rs) numbers yet. The rs numbers will be assigned by the next dbSNP release. (**3/29/06**)

I downloaded 30 submitted SNP (ss) accessions from the SubSNPAcc.bcp.gz file and submitted them as a batch query, but I have yet to receive a fasta file back.

The batch query does not work for unclustered SNPs, and since the records you submitted represent submitted SNPs (ss) that were not mapped and/or clustered, you did not receive a response. You can download the records for these SNPs from the dbSNP FTP site [ftp.ncbi.nih.gov/snp/organisms/macaque_9544/ss_fasta/2005].

(**7/18/06**)